Evaluierung eines Text-Mining-Systems zur Dokumentklassifizierung für das Patentinformationssystem der DaimlerChrysler AG

Diplomarbeit
im Fach Informationstechnik
Studiengang Informationswirtschaft
der
Fachhochschule Stuttgart -
Hochschule der Medien

Petra Klamer

Erstprüfer:	Prof. Dr. Wolf-Fritz Riekert
Zweitprüfer:	Dr. Eberhard Heinz

Bearbeitungszeitraum: 02. April 2004 bis 02. August 2004

Stuttgart, Juli 2004

Kurzfassung

Um eine vereinfachte Patentklassifikation der DaimlerChrysler-Patentschutzrechte zu ermöglichen, wurde für das Patentinformationssystem der DaimlerChrysler AG im Rahmen einer Evaluierung ein Text-Mining-Verfahren auf seine Funktionalitäten in Bezug auf seine Nützlichkeit untersucht und bewertet.

Die vorliegende Arbeit schildert dabei die Analyse des Verfahrens nach festgelegten Kriterien. Diese Qualitätskriterien werden beschrieben und zu ihrer Erfüllung konstruktive Maßnahmen dargestellt und angewandt. Für einen zukünftigen Einsatz innerhalb von DaimlerChrysler wird eine Empfehlung ausgesprochen.

Die internen Patentschutzrechte, die nach den automobilrelevanten Kerntechnologiefeldern innerhalb des DaimlerChrysler Konzerns klassifiziert sind, werden um die Schutzrechte der Wettbewerber ergänzt. Das Text-Mining-Verfahren ermöglicht dabei eine halbautomatische Klassifikation und erleichtert den derzeitigen Zuordnungsprozess immens. Gleichzeitig soll das Verfahren für die Zukunft den Prozess komplett ablösen und vollständig in die täglichen Arbeitsabläufe integriert werden.

Des Weiteren ist durch das Textklassifikationssystem eine Recherche möglich, mittels derer die internen und externen Patentschutzrechte direkt angesehen werden können. Das Verfahren, das auf einer mathematischen Arbeitsweise beruht, fahndet dabei nach ähnlichen Texten und sorgt somit für einen schnellen Überblick über das jeweilige Themengebiet.

Schlagwörter: Ähnlichkeitsberechnung, Evaluierung, Information Retrieval, Patentinformation, Patentklassifikation, Text-Mining

Abstract

In order to allow a simplified classification of the DaimlerChrysler patent protection rights, a Text Mining Method was examined in the course of an evaluation and evaluated in terms of its functionalities with respect to the usefulness for the Patent Information System of the DaimlerChrysler AG.

The present elaboration describes the analysis of the Method according to pre-defined criteria. These quality criteria are described, and in order to meet them, constructive measures are presented and applied. A recommendation is expressed for a future use within DaimlerChrysler.

The classification procedure of the proprietary patent protection rights of the DaimlerChrysler Group shall also be applied on the protection rights of the competitors. The Text Mining Method allows a semi-automatic classification and simplifies the present process of allocation significantly. For the future, it is also contemplated to replace the present process of allocation by the Text Mining Method and to completely integrate the Method into the daily working procedures.

Furthermore, the text classification system allows patent investigations, thus providing a useful basis for direct comparisons between proprietary and external patents.

The Method, which relies on the basis of a mathematical procedure, searches for similar texts and provides for a fast overview of the respective topic.

Keywords: computation of similarity measures, evaluation, information retrieval, patent information, patent classification, text mining

Inhaltsverzeichnis

Abbildungsverzeichnis

Tabellenverzeichnis

Abkürzungsverzeichnis

HdM	Hochschule der Medien
IPM/C	Intellectual Property Management/Central Functions
WR	What's Related
IPC	International Patent Classification
USPOC	U.S. Patent Office Classification
PCT	Patent Cooperation Treaty - Patentzusammenarbeitsvertrag
PARS	Patent Archive and Retrieval System
WIPO	World Intellectual Property Organization
EPA	Europäisches Patentamt
DPMA	Deutsches Patent- und Markenamt
USPTO	US Patent and Trademark Office
JAPIO	Japanese Patent Information Organisation
BRS	Bibliographic Retrieval System
TS	Technologieschlüssel
SVM	Support-Vector-Machine
PAN	Primary Accession Number
MS	Microsoft

Vorwort

People would rather live with a problem they can not solve than accept a solution they cannot understand.

Woolsey and Swanson

Diese Arbeit entstand im Frühjahr/Sommer 2004 als Diplomarbeit im Rahmen der Ausbildung zur Diplom-Informationswirtin im Studienfach Informationswirtschaft an der Fachhochschule Stuttgart, Hochschule der Medien (HdM). Hierbei hatte die Autorin die Gelegenheit, für die Patentabteilung (Intellectual Property Management/Central Functions, kurz IPM/C) der DaimlerChrysler AG eine Software-Komponente zu evaluieren und eine Entscheidungshilfe für einen eventuellen Einsatz zu erarbeiten. Das Aufgabenspektrum umfasste dabei ausgehend von der Erfassung des Ist-Zustands die Konzeption der Testvorgänge und das Austesten der Funktionalitäten, woraus schließlich eine Entscheidung für eine eventuelle Investition in das Produkt zustande kommen sollte.

An dieser Stelle möchte ich mich recht herzlich bei all denen bedanken, die mir diese interessante Aufgabe ermöglicht haben und mir bei der Erstellung der Arbeit stets zur Seite standen. Mein besonderer Dank gilt hierbei Herrn Dr. Eberhard Heinz (Leiter des Teams Information und Dokumentation innerhalb der Abteilung IPM/C), der mir den Ansporn zu dem Thema und der Arbeit gab und für meine Fragen stets ein offenes Ohr hatte, bei Herrn Prof. Dr. Wolf-Fritz Riekert, meinem Diplomarbeitsbetreuer, der trotz seines Studiensemesters engagiert die Betreuung übernahm, allen weiteren Mitarbeitern der Patentabteilung, die mich tatkräftig unterstützten und schließlich den fleißigen Korrekturlesern Herrn Wolfgang Günther, Herrn Werner Fleischer und Frau Tina Slanec. Des Weiteren möchte ich meinen Dank an Herrn Thomas Schmidt und meine Familie aussprechen, die mich während der Zeit in allen Belangen unterstützt haben.

Über entsprechende Rückmeldungen (konstruktive Kritik, Verbesserungsvorschläge, oder sonstige Meinungen) an meine eMail-Adresse (petraklamer@gmx.de) würde ich mich freuen.

1 Einleitung

1.1 Überblick

In Zeiten der digitalen Informations- und Dokumentenflut gewinnen leistungsfähige Such- und Sortierverfahren immer mehr an Bedeutung. Da die Menge an Text in jeder Organisation, Institution oder Firma explosionsartig wächst, ist es mittlerweile von Nöten, dass sich relevante Informationen zu einem bestimmten Thema schnell herauspicken oder sich Dokumente ähnlichen Inhalts zuverlässig gruppieren lassen. Für diese Aufgabe gibt es zwar bereits eine Reihe von Suchmaschinen in Intranet oder Internet, doch deren Trefferquote begeistern laut Ingrid Renz[1] nicht unbedingt alle Benutzer und schon gar nicht die Forschung von DaimlerChrysler. Vor allem in einem forschenden Automobilunternehmen nimmt die Patentinformation eine zentrale Stellung für den Informationsbedarf ein - sie fungiert als ein geeignetes Instrumentarium für einen relevanten Wissensbezug. Aber auch zum Aufbau von Wissensvorsprüngen und Wettbewerbsvorteilen eignet sich die Analyse der Patentanmeldungen allemal, da in ihnen 85 bis 90 Prozent des gesamten veröffentlichten technischen Wissens gespeichert ist. Angesichts der Tatsache, dass jährlich weltweit mehr als eine Million Patentdokumente dem Informationspool der Patentämter neu hinzugefügt werden, werden die Grenzen der traditionellen, durch den Benutzer durchgeführten Analysen schnell deutlich.

Um die vorhandene Informationsflut zu kanalisieren und Benutzern einen bedarfsgerechten Zugriff zu ermöglichen, müssen deshalb neue Wege beschritten werden. Dabei bedarf es nicht nur neuer intelligenter Verfahren für die Suche in unstrukturierten Daten wie Text, Bilder und Audiodaten - das sind etwa 90 Prozent in Unternehmen - sondern auch anderer Ansätze zur Aufbereitung und Strukturierung der unternehmenseigenen Patente, zur Wettbewerbsanalyse sowie zur Analyse von technischen Entwicklungen.
Das rasante, auf allen Technologiegebieten zunehmende Wissen sollte daher so kanalisiert werden, dass es vielen Nutzern bereitgestellt werden kann. In einer Zeit schneller Technologieentwicklungen ist es für ein Unternehmen deshalb wichtig, ständig „up to date“ zu sein, um wichtige Entscheidungen in der Forschung, Entwicklung und Vermarktung neuer Produkte treffen zu können. Ebenso wichtig ist es, den aktuellen Stand der Technik sowie internationales Markt- und Konkurrenzverhalten zu kennen.

Auf Grund dessen wurde nach einer wirkungsvollen Möglichkeit gesucht – und herauskam Text-Mining, gewissermaßen *„das Schürfen nach intellektuellem Gold“*[2], dessen Ziel es ist, aus großen Datenmengen die relevanten Informationen, also das „Wissen“, zu extrahieren.[3]

---

[1] DaimlerChrysler AG (2002), S. 36
[2] DaimlerChrysler AG (2002), S. 36
[3] Vgl. Runkler (2000), S. V

Text-Mining identifiziert und analysiert nach Dirk Krause[4] aus halb- oder unstrukturierten Textdatenbeständen die für den Nutzer interessanten Informationen. *„Dieser Vorgang ist dabei sehr komplex und erfordert aufwändige Methoden, um aus den Dokumenten zusammenhängende Kerninformationen zu extrahieren bzw. weiterzuverarbeiten."*

Unter dem Begriff des Text-Mining wird dabei eine Sammlung von Methoden, Technologien und Produkten verstanden, die in unterschiedlichen Zusammenhängen von Bedeutung sind.

Die Techniken selber erlauben vor allem eine automatische, schnelle und Domänenübergreifende Indexierung von Dokumenten, was zu einer effizienteren Verwaltung führen kann. Ebenso bezeichnet Text-Mining die Extraktion von Wissen aus nicht oder nur teilweise strukturierten Daten. Dieses Wissen gilt es dabei nutzbar zu machen und die darin steckenden Potentiale für Anwendungen zu erkennen. So können die Datenbestände unter anderem daraufhin untersucht werden, ob sie das gleiche oder nur ein verwandtes Thema behandeln.[5]

Wo das nicht genügt, ist man bei der Suche nach weiterführenden Informationen laut Thomas Kamphusmann[6] *„auf die Experten angewiesen, die Patente, nach denen man sucht, kennen – bis man selber einer dieser Experten ist."* Damit ist die Problemlage angedeutet: Je spezifischer die Informationsbedürfnisse sind, desto weiter muss in die Informationsangebote eingetaucht werden, um beurteilen zu können, ob das jeweilige Informationsbedürfnis überhaupt befriedigt wird.
Zur Anwendung kommen im Text-Mining dabei verschiedene Methoden und Hilfsmittel wie statistische und linguistische Verfahren. Statistische Verfahren spielen bei der Strukturierung von Texten eine grundlegende Rolle und werden unter anderem für die Dokumentenindizierung eingesetzt. Diese Verfahren versuchen nicht, die tiefer liegende Bedeutung eines Wortes zu ermitteln, sondern dienen als semantische Indikatoren. Es werden Maßzahlen der Häufigkeit des Auftretens eines Terms innerhalb eines Dokuments bzw. in einer Dokumentensammlung als Anhaltspunkt für eine geringere und höhere Bedeutung hinsichtlich des Inhalts gesehen. Linguistische Verfahren basieren auf der algorithmischen Beschreibung einer Sprache. Die Analyse eines Textes erfolgt aus linguistischer Sicht in verschiedenen Ebenen der Textrepräsentation:

- Morphologische Ebene

  Hier wird die Struktur von Wörtern bestimmt. Der Ansatz versucht, Terme nicht als Zeichenketten zu definieren, sondern als bestimmte Form eines Wortes aufzufassen. Dabei wird zwischen Grundform- und Stammformreduktion unterschieden. Die Grundformreduktion führt Wörter auf ihre grammatikalische Grundform zurück, die Stammformreduktion extrahiert aus den Wortformen den zugehörigen

[4] Krause (2002), S. 581
[5] Vgl. DaimlerChrysler AG (2002), S. 36
[6] Kamphusmann (2002), S. 13-14

Stamm, der im Allgemeinen keine in der Sprache als Wort vorkommende Form ist. Diese Reduktionen werden auch als Lemmatisierungen bezeichnet.

- Lexikalische Ebene

  Hier werden einzelne Wörter untersucht. Einfache Verfahren sind in der Lage, Texte effizient in die enthaltenen Wörter zu zerlegen. Indexierungsterme als Wortfolgen repräsentieren Texte als Zeichenketten in Abfolgen von Wörtern. Die Reihenfolge der Wörter wird allerdings vernachlässigt. Dokumente werden also nicht mehr als Sequenzen von Wörtern dargestellt, sondern als unterschiedliche Menge von Wörtern. Jedes Wort ist ein Attribut mit unterschiedlichen Werten. Die Werte selber bilden sich aus der Anzahl der Häufigkeit eines entsprechenden Wortes. Sie wird auch als Term-Frequenz bezeichnet.

- Syntaktische Ebene

  Hier wird die Struktur der Sätze untersucht. Die Idee besteht darin, Indexierungsterme nicht nur aus einzelnen, sondern aus mehreren Wörtern bestehen zu lassen. Somit werden Sätze in ihre Bestandteile zerlegt und die Beziehungen zwischen den Wörtern analysiert.

- Semantische Ebene

  Wenn die Semantik von Dokumenten erfasst werden könnte, würde ein Text-Mining-Algorithmus optimal arbeiten. Mit dem derzeitigen Stand der Forschung können allerdings nur eingeschränkte Aussagen über die Bedeutung der Dokumente mit Hilfe von statistischen Analysen getroffen werden. In dieser Ebene werden unter anderem semantische Kategorien gebildet, z.B. mit der Methode des Term-Clustering.

- Pragmatische Ebene

  Diese Ebene erweitert die semantische, indem die Bedeutung des Textes erweitert wird.

1.2 Aufbau der Diplomarbeit

Kapitel 1 gibt einen kurzen Einblick in das Themengebiet der Diplomarbeit.

In **Kapitel 2** wird die vorgegebene Aufgabenstellung aufgezeigt. Daraus ergeben sich die Zielsetzungen dieser Arbeit. Auch werden in dem Kapitel genaue Begriffsdefinitionen festgelegt.

Das **Kapitel 3** beschäftigt sich mit der inhaltlichen Ausgangslage. Es gibt einen Überblick über die aktuelle Situation innerhalb der Patentabteilung der DaimlerChrysler AG und die derzeitige Aufgabenbewältigung. Des Weiteren wird auf die bestehende Problemsituation aufmerksam gemacht. Darauf aufbauend wird das Kapitel an der Stelle durch den direkten Bezug an die gestellten Anforderungen an ein Text-Mining-Verfahren ergänzt.

Auf die thematisch verwandten Gebiete des Text-Mining und die im Text-Mining verwendeten Techniken wird in **Kapitel 4** genauer eingegangen.

Das **Kapitel 5** beschäftigt sich mit der Analyse des angestrebten Verfahrens. Dazu zählt die Arbeitsweise des Text-Mining-Systems sowie dessen Anwendung durch den Benutzer.

Kapitel 6 dagegen geht zu Beginn auf die theoretische Erläuterung der Evaluierung des Verfahrens ein. Dieses Kapitel wird erweitert durch die realisierten Untersuchungen und quantitativen Messungen.

Das Ergebnis davon wird in **Kapitel 7** beschrieben. Die tatsächlich erreichten Ziele und der Nutzen des Systems bilden einen wichtigen Bestandteil. Darauf aufbauend wird eine Empfehlung für oder gegen das System abgegeben. Ferner wird ein Ausblick auf eine mögliche Erweiterung und zukünftige Nutzungsmöglichkeit gegeben.

Kapitel 8 gibt eine Zusammenfassung der durchgeführten Arbeit.

Der **Anhang** enthält schließlich weitere Ergebnistabellen zu Kapitel 6 sowie das **Literaturverzeichnis**.

2 Aufgabenstellung und Zielsetzung

Anhand der vorgegebenen Aufgabenstellung werden die Zielsetzung und notwendige Definitionen dieser Arbeit festgelegt.

2.1 Aufgabenstellung

Die gestellte Aufgabe lautet, das Text-Mining-Verfahren „What's Related" (WR), das von der konzerninternen Forschungseinheit „Department of Data Mining Solutions, RIC/AM" als Basisgerüst entwickelt und im Rahmen des Projekts an die Bedürfnisse der Patentabteilung angepasst wurde, zu evaluieren und seine Funktionalität nach ausgewählten Kriterien zu bewerten.

Nach dem Brockhaus[7] ist Evaluierung (Synonym Evaluation)

> *„die Analyse und Bewertung eines Sachverhalts, vor allem als Begleitforschung einer Innovation. […] quasi zur Überprüfung der Eignung eines sich in Erprobung befindlichen Modells. Evaluation wird auch auf die Planung angewendet, zum Zweck der Beurteilung der Schlüssigkeit der Zielvorstellung und der zu deren Verwirklichung beabsichtigten Maßnahmen. Bei der Analyse eines gegebenen Faktums ist Evaluation die Einschätzung der Wirkungsweise, Wirksamkeit und der Wirkungszusammenhänge*.[…]"

Das System soll dabei im Endeffekt zwei Funktionalitäten gewährleisten:

1. Klassifikation von Patentschutzrechten gemäß einer innerhalb des Konzerns bestehenden Klassifikationsvorgabe und anschließende Speicherung in einer Datenbank. Mit den Daten der Datenbank können im weiteren Gebrauch unter anderem Statistiken erstellt werden.
2. Über eine Oberfläche soll eine Recherche nach thematisch ähnlichen Patenten möglich sein. Die Recherche greift dazu ebenfalls auf die in Punkt 1 genannte Datenbank zu.

2.2 Zielsetzung

Durch die Evaluierung soll geklärt werden, ob eine Investition in das System überhaupt zu rechtfertigen ist. Geklärt werden soll, ob die gegebenen Funktionalitäten die erwarteten Anforderungen von IPM/C abdecken und ob die Software die Anforderungen bzw. die an sie gestellten Aufgaben erfüllen kann.

Dies beinhaltet sowohl den Umfang, das heißt die Anzahl und Vollständigkeit, der darin enthaltenen Patentdokumente, als auch deren Auffindbarkeit (retrieval capability).

---

[7] Brockhaus (1997), S. 716

Diese sagt aus, ob und wie diese Dokumente bei einer Recherche gefunden werden können, was natürlich unmittelbar im Zusammenhang mit der Qualität der inhaltlichen Erschließung der Patentschutzrechte steht. Der Gesichtspunkt der Handhabbarkeit für den Endbenutzer aus Forschung und Entwicklung oder den Informationsspezialisten wird dabei nicht betrachtet.

Durch die Evaluierung soll letztendlich eine Situationsverbesserung erfolgen. Verbesserung bedeutet unter diesen Umständen, dass die derzeitige Form der Aufgabenbewältigung verändert und eventuell verbessert werden kann. Genauer gesagt wird mit der Untersuchung überprüft, ob das Text-Mining-Verfahren dem Anwender eine Erleichterung bei der Ähnlichkeitssuche verschafft, ohne dabei ungenaue oder falsche Aussagen zum Inhalt der Patentschutzrechte zu erhalten, und ob eine einfachere und verbesserte Klassifikation erfolgen kann.

Im Falle eines positiven Testergebnisses soll das System weiter ausgebaut, die entsprechenden Oberflächen realisiert und schließlich als letzte Hürde die Integration des Systems innerhalb der DaimlerChrysler AG durchgeführt werden. Werden die Anforderungen hingegen nicht erfüllt, muss nach weiteren Alternativen Ausschau gehalten werden.

2.3 Festlegungen

Um die Diplomarbeit auf einer einheitlichen Basis aufzubauen, werden im Folgenden verschiedene Definitionen festgelegt.

- Um das Lesen der Arbeit zu erleichtern, wird auf Doppelbezeichnungen (z.B. Informationsspezialist/ Informationsspezialistin) verzichtet und nur die männliche Form verwendet.
- Ebenfalls um der besseren Lesbarkeit willen, wird auf Kenntlichmachung geschützter Firmen- bzw. Produktnamen verzichtet. Es ist deshalb generell davon auszugehen, dass alle in dieser Arbeit erwähnten derartigen Namen geschützt sind.
- Wenn im weiteren Verlauf der Einfachheit halber von Patenten, Patentschutzrechten oder Patentdokumenten die Rede ist, sind damit die Patentanmeldungen sowie die erteilten Patente gemeint.

3 Das Projekt

Dieses Kapitel gibt einen Überblick über die inhaltliche Ausgangslage der Arbeit. Nach einem Einblick in das Themengebiet des Patentwesens erfolgt die Beschreibung der aktuellen Situation innerhalb der Patentabteilung IPM/C sowie des derzeitigen Realisierungsprozesses der Klassifikation.

3.1 Wissensgrundlagen

Damit für die vorliegende Arbeit eine grobe Wissensbasis geschaffen wird, soll das Patentwesen in den folgenden Unterkapiteln genauer erläutert werden. Im Rahmen der weiteren Evaluierung ist dies unumgänglich, da das gewonnene Wissen ein besseres Verständnis der durchgeführten Untersuchungen in Kapitel 6 ermöglicht.

3.1.1 Patentschutzrechte

Der Schutz von Eigentum ist in unserem Rechtssystem verankert. Nicht nur materielle Werte, sondern auch Ideen und intellektuelle Leistungen werden geschützt. Für geistiges Eigentum technischen Charakters hat sich der Schutz durch Patente[8] und Gebrauchsmuster[9] entwickelt. Vereinfacht dargestellt bekommt nur derjenige Schutz für eine Erfindung, der diese zum Patent bzw. Gebrauchsmuster anmeldet. Die maximale Laufzeit von Patenten beträgt im Allgemeinen 20 Jahre ab Anmeldetag.[10]

Patente müssen dabei bei eigens dafür zuständigen Behörden angemeldet werden. Dafür gibt es bestimmte Formvorschriften, Gebühren sind zu entrichten und festgelegte Fristen einzuhalten. Nach erfolgter Anmeldung der Erfindung spricht man von einer Patentanmeldung. Diese wird bei einer Erstanmeldung nach Ablauf einer Frist von 18 Monaten offen gelegt bzw. veröffentlicht.[11] Dazu werden von den Patentämtern die Druckschriften - die den Text der Anmeldungen enthalten - herausgegeben. Diese Schriften nennen sich Offenlegungsschriften. Da der Patentschutz stets auf ein Land bezogen ist, muss dieselbe Erfindung bei jedem Zielland angemeldet werden, was die

---

[8] Patent nach net-Lexikon.de (2004):

Das Patent ist ein gewerbliches Schutzrecht, das ein ausschließliches Recht zur gewerblichen Nutzung eines technischen Verfahrens oder eines technischen Produkts gewährt. Es wird demjenigen verliehen, der ein bestimmtes Verfahren zuerst in einer Patentschrift veröffentlicht. Es ist zeitlich begrenzt und schützt vor unbefugter Nachahmung.

[9] Gebrauchsmuster nach net-Lexikon.de (2004):

Das Gebrauchsmuster ist ebenfalls ein technisches Schutzrecht, welches auch häufig als das „kleine Patent“ bezeichnet wird. Hierbei gibt es allerdings keinen Schutz für Verfahren, sondern nur für Erfindungen. Auch beträgt die Schutzfrist nur 3 Jahre und ist auf maximal 10 Jahre zu verlängern. Ein Gebrauchsmuster ist jedoch schneller, günstiger und einfacher zu erlangen. Patent und Gebrauchsmuster können auch nebeneinander bestehen.

[10] Vgl. Bendl/Weber (2002), S. 1

[11] Vgl. Bendl/Weber (2002), S. 4

Anzahl der technisch identischen Dokumente weiter hochtreibt. Deshalb erfolgt die Zusammenführung der Dokumente zu einer Patentfamilie.

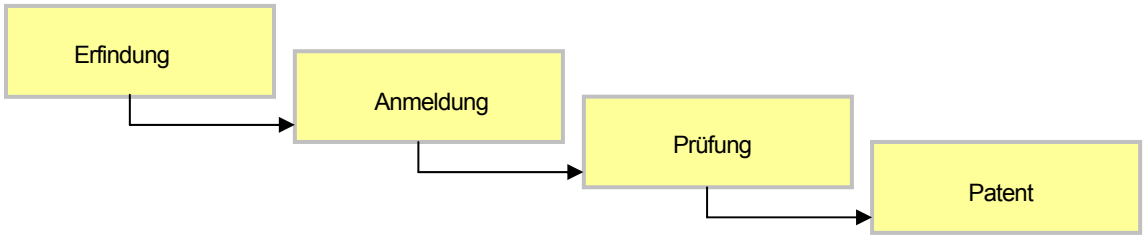


Abbildung 1: Von der Erfindung zum Patent[12]

Um zu ermitteln, was auf dem entsprechenden Gebiet der Erfindung schon bekannt ist, wird zunächst eine Recherche zum Stand der Technik durchgeführt. Eine derartige Recherche ist wichtig, da einerseits eine Patentanmeldung viel Geld kostet und andererseits natürlich nichts angemeldet werden sollte, was es schon gibt. Gleichfalls können die bei einer Recherche gefundenen Dokumente zu neuen Erkenntnissen führen und in weiteren oder besseren Erfindungen münden.

Eine Patentschrift oder Patentanmeldung besteht dabei aus mehreren Teilen:

1. **Titelseite**: Hier sind die bibliographischen Daten der Patentanmeldung zu finden. Dazu zählen die Publikationsnummer, die Anmeldenummer, der Erfinder, eine Zusammenfassung und ein eventuelles Prioritätsdatum.
2. **Beschreibung**
3. **Ansprüche**
4. **Zeichnungen**

Die Patentschriften selber stehen in Patentdatenbanken, die ihrerseits nach der Art der zur Verfügung gestellten Informationen unterteilt werden. Zu unterscheiden sind:

- ❖ Bibliographische Datenbanken

Die Suchfelder beinhalten im Allgemeinen den Titel der Anmeldung, den Anmelder, die Anmeldedaten, den Erfinder, die Familienmitglieder, die Prioritätsangaben und die Patentklassen. Weitergehende Informationen, die den Inhalt der Patentanmeldung betreffen, sind nicht vorhanden.

- ❖ Abstract-Datenbanken

Hier steht zusätzlich eine Kurzfassung (Abstract) der Anmeldung zur Verfügung.

- ❖ Volltextdatenbanken

Der gesamte Text der Anmeldung ist in der Datenbank vorhanden und kann durchsucht werden. Diese Art der Datenbank eignet sich besonders zur Suche nach

[12] Quelle: Eigene Darstellung

bestimmten Textstellen. Allerdings ist bei dieser Datenbank problematisch, dass die Recherche oft eine übergroße Anzahl an Treffern ergibt, die mit dem gesuchten Ergebnis wenig bis gar nichts zu tun haben. Das geschieht, weil das Suchprogramm nur einzelne Suchbegriffe auffindet. Diese Art von Datenbank gehört aber mit zu den effektivsten Recherchewerkzeugen.

- ❖ Faksimile-Datenbanken

Darunter wird eine Sammlung von Abbildungen der originalen Patent- oder Anmeldeschriften verstanden. Es handelt sich dabei um fotografische Abbilder der Originalschriften, bestehend aus Titelblatt, Beschreibung, Ansprüchen, Zeichnungen und eventuellen Rechercheberichten.[13]

3.1.2 Warum Patentinformation?

Nach Lothar Wild und Alfred Wittmann[14] werden unter dem Begriff „Patentinformation" zwei Dinge verstanden: Zum einen handelt es sich um die Unterrichtung über neueste Ergebnisse aus Entwicklung und Forschung in allen technischen Gebieten, und zum anderen geht es um die Unterrichtung über Existenz und Umfang bereits bestehender Patentschutzrechte. In der Planung dienen Patentinformationen der unternehmerischen Entscheidungsfindung, gleichzeitig ermöglicht die Transparenz des Patentwesens eine realistischere Einschätzung des Standorts der eigenen Wirtschaft. Die Möglichkeit der Unterrichtung über die Existenz von Patentschutzrechten dient einerseits dem Schutz der Erfindung, andererseits soll sie den Unternehmen helfen, Entwicklungen, die bereits in geschützter Form vorliegen, nicht nochmals in Angriff zu nehmen und somit Doppel- oder Fehlinvestitionen zu verhindern. *„Allein in Deutschland werden so jedes Jahr [...] etwa 12,5 Milliarden Euro für unnötige Doppelforschung und -entwicklung vergeudet"*, erklärt dazu Alexander Wurzer[15].

Deutlich wird damit allemal, dass Patente zwei Funktionen innehaben: die Schutzrechts- und die Informationsfunktion. Sie fungieren demnach als eine Möglichkeit, Forschung und Entwicklung messbar und vergleichbar zu machen.[16] Da die in Patentanmeldungen dokumentierten Technologien zum Zeitpunkt der Anmeldung und ihrer Veröffentlichung üblicherweise noch nicht zum Einsatz gelangt sind, eignen sich Patentdaten in besonderer Weise gleichfalls als Frühindikatoren für bevorstehende technische Entwicklungen. Auch gilt es zu ermitteln, welche Positionen in einer Branche unter Umständen prinzipiell erfolgsträchtiger sind als die eigene. Daraus kann analysiert werden, wie man als Unternehmen selber reagieren muss - ob mehr in bestimmte Bereiche investiert werden sollte oder ob bestimmte Bereiche schon sehr gut dastehen. Diese Stufe sollte natürlich für andere Wettbewerber nicht erreichbar sein.[17]

---

[13] Vgl. Bendl/Weber (2002), S. 35-40
[14] Wild/Wittman (1990), S. 12-17
[15] Wurzer (2003), S. 38
[16] Vgl. Hofmann (2000), S. 16-17
[17] Vgl. Bartölke (2000), S. 7-8

Darüber hinaus kann aus den spezifischen Bewegungen der Patentanmeldungen entnommen werden, welcher Natur eine Entwicklung ist, ob es sich um einen von kleinen Schritten getragenen kontinuierlichen Prozess handelt, ob in einem bestimmten Bereich größere technologische Durchbrüche gelungen sind oder ob sich ganz neue technologische Felder eröffnen. Auch das Beziehungsmuster von Patenten untereinander kann aufschlussreich sein. Gedacht wird hierbei an die Beobachtung, dass eine Basiserfindung regelmäßig einen Schwarm von Folgeerfindungen nach sich zieht.[18]

Zusammenfassend werden deshalb nach Heinrich Fendt[19] mittels Patentinformationen Antworten auf folgende typische Fragestellungen gesucht:

- Welche Unternehmen sind Technologieführer?
- Womit beschäftigen sich Branchenneulinge?
- Welches sind die wichtigsten Absatzmärkte für technische Lösungen?

Oder es werden Untersuchungen angestellt über:

- Patentaktivitäten von Wettbewerbern und Ländern im Rahmen der Beurteilung eigener Forschungs- und Entwicklungs-Anstrengungen sowie der Innovationskraft und Stellung im Wettbewerb
- Identifizierungen von offensiv und defensiv agierenden Unternehmen und Ländern

Aber auch die eigene Forschung und Entwicklung kann gemessen werden:

- Prüfen, Justieren und Definieren von Forschungsrichtungen unter Berücksichtigung wichtiger Basispatente und Wettbewerberstrategien.
- Offenlegen der Vernetzungen von Erfindungen und Technologiefeldern sowie deren Berührungspunkte zu Nachbardisziplinen.

Laut Alexander Wurzer[20] werden in dem Zusammenhang immerhin 85 bis 90 Prozent des technischen Wissens in der Patentliteratur publiziert. Dabei werden nur etwa 10 bis 15 Prozent dieses veröffentlichten Wissens in der sonstigen Literatur wiedergegeben. Und das kann sogar erst bis zu fünf Jahre nach der Anmeldung zum Patent erfolgen. Damit ist die Patentliteratur die wichtigste technische Informationsquelle überhaupt.

Die Bedeutung der technischen Informationen wird weiterhin durch die enorme Anzahl von technischen Patentschutzrechten verdeutlicht. Gegenwärtig sind vier bis fünf Millionen Patente und Gebrauchsmuster weltweit in Kraft, und es kommen jährlich etwa eine halbe Million neuer Patentschutzrechte hinzu. Sie dokumentieren das technische Wissen in einzigartiger Weise und umfassen dabei sämtliche Gebiete der Technik in allen wichtigen Regionen der Erde.

---

[18] Vgl. Greif/Potkowik (1990), S. 3-4

[19] Fendt (2004)

[20] Wurzer (2003), S. 27

Da Patentanmeldungen in aller Regel nach achtzehn Monaten offen gelegt werden, ist diese Art von Informationsquelle ausgesprochen aktuell. Im Gegensatz dazu ziehen sich Produktentwicklungen oft über Jahre hin. Innerhalb der ersten ein bis zwei Jahre der Produktentwicklung gibt es außer der Veröffentlichung der Patentanmeldung 18 Monate nach der Einreichung der Patentanmeldung in der Regel keine andere Informationsquelle. Erst sehr viel später, wenn die Vorprodukte schon getestet und erprobt werden, erscheinen weitere Informationen über die Produkte.[21]

Elektronische Aufbereitung von Patentdokumenten und die Bereitstellung in Datenbanken bilden deshalb die Voraussetzung für einen automatisierten Zugriff und eine systematische Einbeziehung von Patentinformationen im Rahmen der betrieblichen Entscheidungsprozesse. Wichtige Voraussetzung dafür ist allerdings ein funktionierendes Informationssystem.

3.1.3 Internationale Patentklassifikation

Patentinformationen stehen im Gegensatz zu den meisten anderen Informationsquellen für Technik und Naturwissenschaft nach einem Ordnungssystem zur Verfügung. Dieses System wird „Patentklassifikation“ oder kurz „Klassifikation“ genannt und ist unter dem Begriff „International Patent Classification“ - kurz IPC - international vereinheitlicht. So ist die gesamte Naturwissenschaft und Technik zu Dokumentations- und Recherchezwecken nach Sachgebieten hierarchisch und sprachunabhängig klassifiziert. Nach Andreas Hofmann[22] bildet „*mit Ausnahme der USA, die neben der IPC hauptsächlich ein eigenes nationales Patentklassifikationssystem (USPOC) benutzt, die IPC in fast allen Ländern das standardisierte Klassifizierungssystem*.“ Sie beinhaltet dabei über 60.000 Sachgebiete. Durch dieses Klassifikationssystem wird die Suche nach ähnlichen Patenten innerhalb eines Fachgebiets erheblich erleichtert.

Die Patentklassifikation ist allerdings kein starres, immer gleich bleibendes System, sondern dynamisch, da das Klassifizierungssystem im Abstand von 5 Jahren revidiert und ergänzt wird. Auf diese Weise soll vor allem der technischen Entwicklung Rechnung getragen werden.

[21] Vgl. Wurzer (2003), S. 35-38
[22] Hofmann (2000), S. 29-30

Das Prinzip besteht darin, dass der gesamte Stand der Technik in Sektionen unterteilt und mit Symbolen versehen ist. Die Sektionen setzen sich dabei folgendermaßen zusammen:

A: Täglicher Lebensbedarf

B: Arbeitsverfahren; Transportieren

C: Chemie; Hüttenwesen

D: Textilien; Papier

E: Bauwesen; Erdbohren; Bergbau

F: Maschinenbau; Beleuchtung; Heizung; Waffen; Sprengen

G: Physik

H: Elektrotechnik

Jede Sektion ist zusätzlich noch in Untersektionen unterteilt, die Klassen, Unterklassen und Gruppen enthalten. Gruppen wiederum bestehen aus Haupt- und Untergruppen. Mit diesem etwas bizarr anmutenden System ist es gelungen, die gesamte Technik einzuteilen und relevante Dokumente wieder auffindbar zu machen.[23]

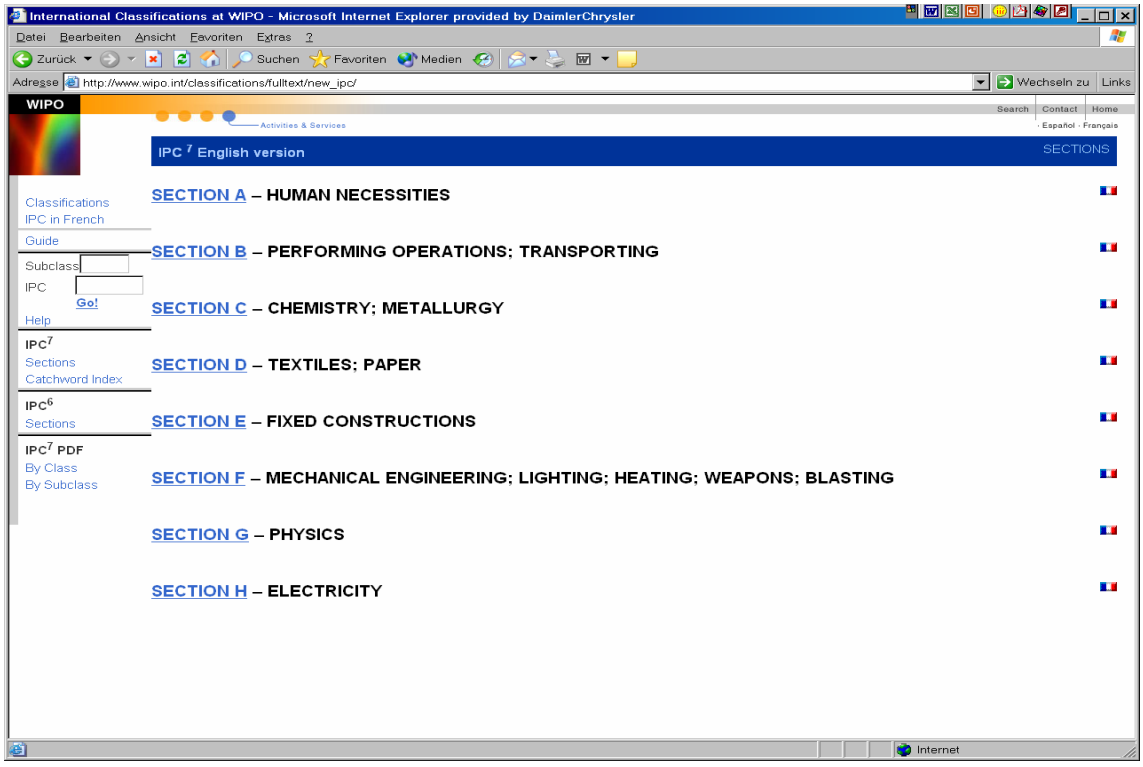


Abbildung 2: Aufteilung der Internationalen Patentklassifikation[24]

[23] Vgl. Bendl/Weber (2002), S. 127

[24] Quelle: World Intellectual Property Organization (2004)

Zur Verdeutlichung ein Beispiel:

Wenn eine Patentanmeldung in Sektion A63C17/02 eingeordnet ist, dann gehört sie in das Gebiet des täglichen Lebensbedarfs zur Klasse Sport/Spiele/Vergnügen.

In dieser Klasse werden (zumindest theoretisch) alle Patentanmeldungen und Patente des oben genannten Themengebiets gefunden. Eine Patentanmeldung, die in A63C17/02 klassifiziert ist, weist alle Eigenschaften der ihr übergeordneten Gruppen und Klassen auf. Allerdings werden die Dokumente nicht in allen höher gestellten Klassen eingeordnet. Dies würde die Anzahl der Dokumente pro Klasse zu groß werden lassen. Dokumente werden in der niedrigstmöglichen Klasse eingeordnet. Der Vorteil liegt auf der Hand: Durch dieses Einteilungssystem ist man von Stich- oder Schlüsselwörtern nicht mehr abhängig. Das heißt, erfasst werden mittels einer Recherche auch unbekannte Synonyme und Dokumente mit Schreib- oder Übersetzungsfehlern.[25]

A	*Sektion*	*Täglicher Lebensbedarf*
63	*Klasse*	*Sport; Spiele; Vergnügen*
C	*Unterklasse*	*Schlittschuhe; Skier; Rollschuhe; Spielplätze*
17	*Gruppe*	*Rollschuhe*
02	*Untergruppe*	*mit zweipaarig angeordneten Rollen*

[25] Vgl. Bendl/Weber (2002), S. 127-129

3.2 Gegenwärtige Situation bei IPM/C

Dieses Kapitel soll einen kurzen Einblick in die Arbeitsumgebung von IPM/C geben. Dazu werden die verwendeten Daten, Verfahren und Systeme genauer erläutert.

3.2.1 Datenbestand

Der Datenbestand der Patentabteilung des DaimlerChrysler-Konzerns ist zweigeteilt.

Derwent	Wila
Nur automobilrelevante Datenprofile nach IPC-Klassen	Sämtliche IPC-Klassen
Anmeldungen der 40 wichtigsten Länder Verfügbare Daten seit 1991	Anmeldungen des Deutschen (seit 1968), des Europäischen (seit 1978), des US- (seit 1975) und des Japanischen Patentamts (seit 1976) sowie der Internationalen Anmeldungen (seit 1978)
Ca. 8 Millionen Veröffentlichungen	Ca. 15 Millionen Patentanmeldungen
Ca. 70 Updates im Jahr	Wöchentliche Updates
Abstracts mit Zeichnungen und Patentfamilie	Bibliographische Daten, Zusammenfassungen und Hauptansprüche

Tabelle 1: Geteilter Datenbestand der Patentabteilung[26]

Durch den britischen Anbieter Derwent werden rund 8 Millionen Veröffentlichungen zu sämtlichen Sachgebieten bereitgestellt und durch etwa 14.000 Einträge wöchentlich ergänzt. Die Datenbank liefert die bibliographischen Angaben zu Patentschriften der 40 wichtigsten Länder, einschließlich der Anmeldungen beim Europäischen Patentamt. Derwent stellt zu den Anmeldungen Angaben über die so genannten Patentfamilien zur Verfügung und weist somit für jede Basisanmeldung aus, zu welchen Auslandsanmeldungen (Äquivalente) sie geführt hat. Der Derwent-Datenbestand wird von der Industrie viel verwendet, da Derwent zu jeder Anmeldung durch technische Fachleute eine eigene Kurzfassung (Abstract) in englischer Sprache erstellen lässt. Diese „intelligenten" Abstracts tragen zur Objektivierung der Sachverhalte bei und verbessern die Suche nach Stichwörtern.[27]

[26] Quelle: Angelehnt an Heinz (2004), Folie 5
[27] Vgl. Fendt (1991)

Wila hingegen liefert als Datenbank die bibliographischen Daten, die Zusammenfassungen und Hauptansprüche[28] des Deutschen, Europäischen, Japanischen und des US-Patentamts sowie der PCT-Anmeldungen.[29] Das Update der Datenbestände erfolgt dabei ebenfalls in wöchentlichen Abständen.

Die Informationsressourcen werden über das Patentinformationssystem PARS (Patent Archive and Retrieval System) bezogen. PARS ist ein umfassendes elektronisches Patentinformationssystem für Großunternehmen, das aus einer Recherchedatenbank und einem Vollschriftenarchiv besteht. Verfügbar sind die Patentschriften der World Intellectual Property Organization (WIPO), des Europäischen Patentamtes (EPA), des Deutschen Patent- und Markenamts (DPMA), des US Patent and Trademark Office (USPTO) und der Japanese Patent Information Organisation (JAPIO).

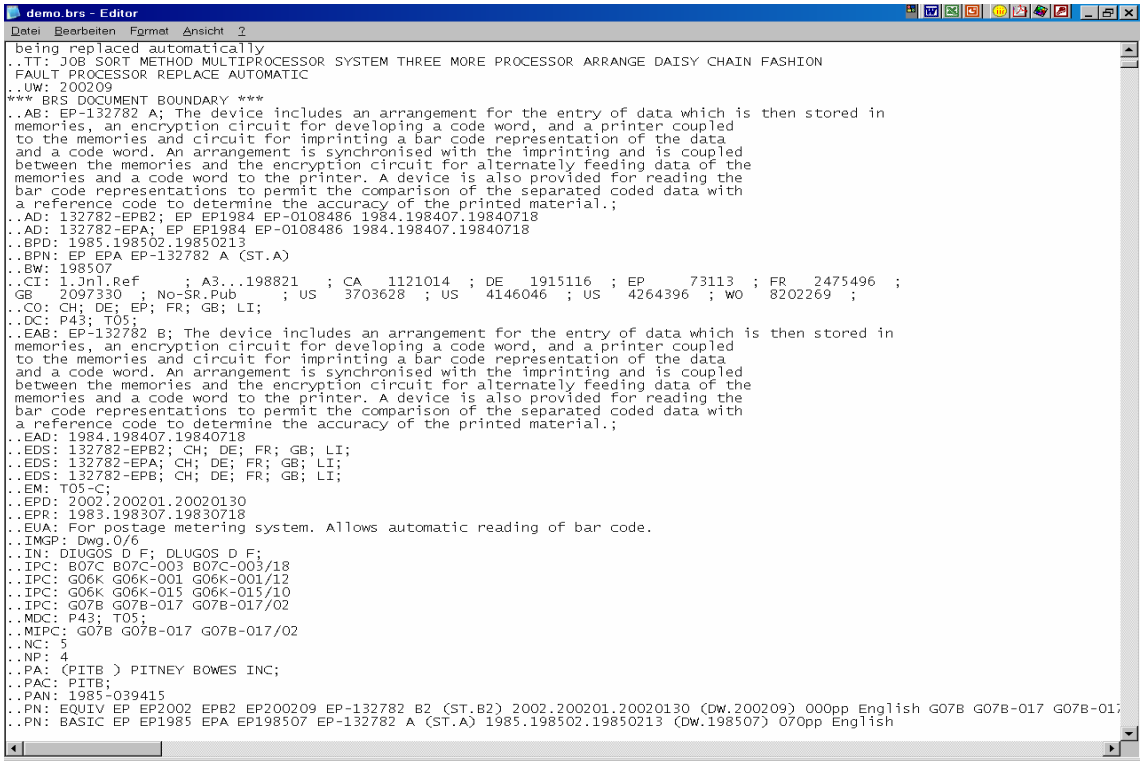


Abbildung 3: Auszug einer Wochenlieferung

Als Ursprungsdatei wird jeweils eine „brs-Datei“ vom Anbieter heruntergeladen und in das bestehende Datenbanksystem eingespielt. Die Datei enthält wie in Abbildung 3 dargestellt jeweils eine komplette Kalenderwoche mit den verschiedenen Patentschriften.

---

[28] Hauptanspruch nach der Fraunhofer-Patentstelle (2004):
Hauptanspruch ist ein unabhängiger Anspruch, der alle wesentlichen Merkmale der Erfindung ohne Bezugnahme auf andere Ansprüche wiedergibt.

[29] PCT-Anmeldung nach Däbritz (1994), S. 34:
Einreichung einer Patentanmeldung (der internationalen Anmeldung) in einer Sprache bei einem nationalen oder regionalen Patentamt (dem Anmeldeamt) mit Wirkung in mehreren Staaten, die der Anmelder in der Anmeldung bestimmt (Bestimmungsstaaten), anstelle der Einreichung mehrerer gesonderter nationaler und/oder regionaler Patentanmeldungen.

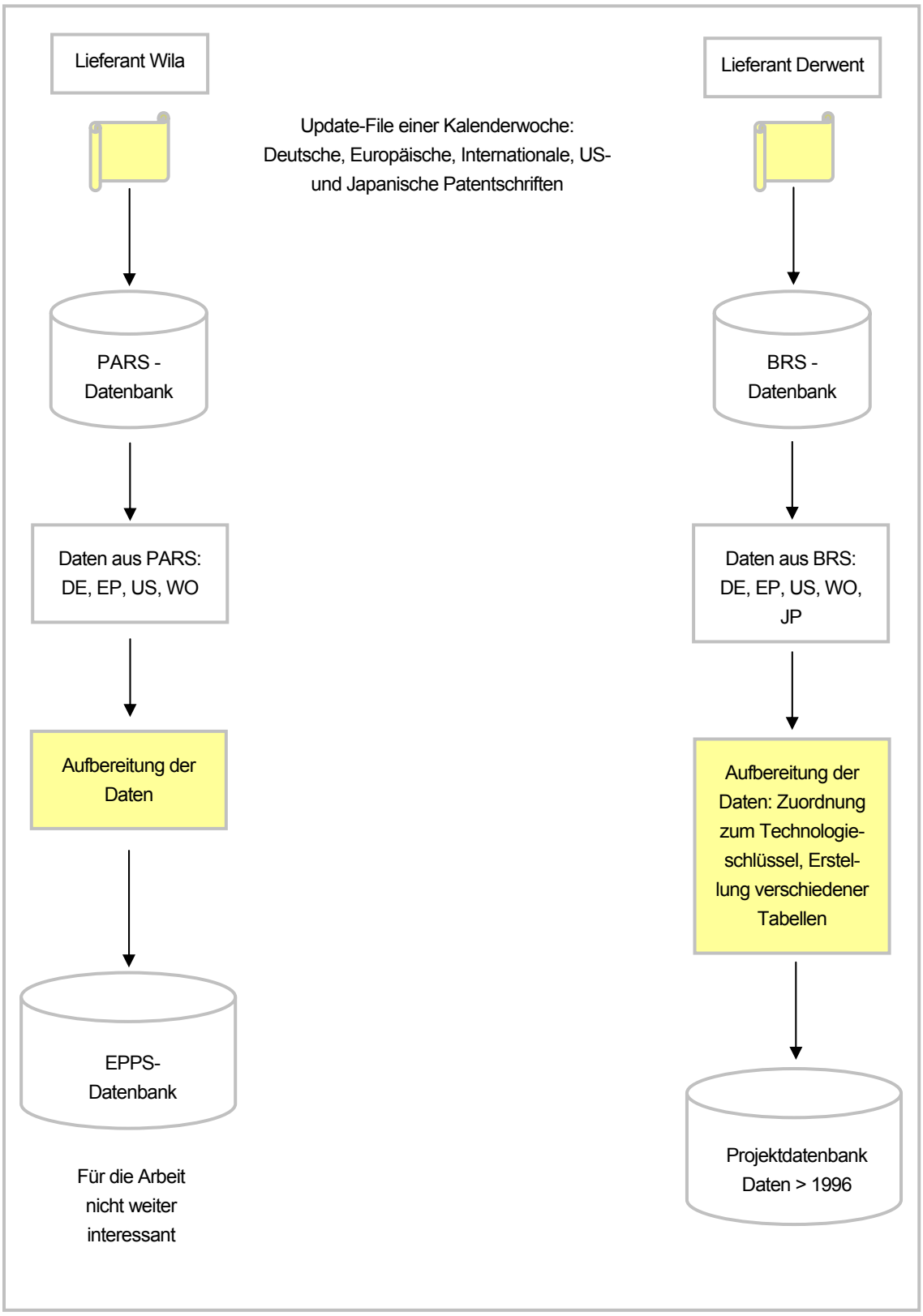


Abbildung 4: Ablauf des Datenbank-Update[30]

Nach Einspielen der Aktualisierungsdatei (oder Update-File) in die beiden Datenbanken werden die Daten unterschiedlich weiterbehandelt.

[30] Quelle: Eigene Darstellung

Ein Teil der Daten - die Deutschen, Europäischen, US- und die Internationalen Anmeldungen - werden aufbereitet und in eine weitere Datenbank geschrieben, wo sie für einen Patentprofildienst[31] genutzt werden. Für die weitere Arbeit ist diese Datenbank jedoch nicht relevant.

Der andere Teil der Daten, mitsamt den Japanischen Schriften, wird ebenfalls aufbereitet und in eine weitere Datenbank geschrieben. Diese Datenbank namens „Projektdatenbank“ beinhaltet alle Daten ab dem Jahr 1996. Den darin enthaltenen Patentschutzrechten wird der so genannte Technologieschlüssel zugeordnet. Die Daten werden dabei mittels der Datenbank-Maschine des BRS-Datenbanksystems verarbeitet. Das Datenbanksystem besteht dabei aus den beiden Komponenten „BRS-Search“ und „Netanswer“.

3.2.2 Technologieschlüssel

Patentschutzrechte beziehen sich nur auf einzelne Erfindungen, ein Fahrzeug hingegen besteht aus vielen verschiedenen Bau- und Funktionsgruppen der unterschiedlichsten Technologien. Deshalb ist hier eine Gruppierung der Patente nach technologischen Kriterien von Vorteil.

Das bedeutet, dass die Internationale Patentklassifikation nur teilweise anwendbar ist. Sie ist für den Automobilbereich zu fein, zu detailliert und hat zum Teil auch eine andere Zielsetzung. Sie orientiert sich, wie bereits erwähnt, an eher wissenschaftlichen Kriterien und nicht allein an den technologischen automobilrelevanten Themen. Somit ist die IPC nicht das optimalste Mittel, um die Patente der DaimlerChrysler AG abzubilden. Ein weiteres Problem besteht zudem darin, dass die Sprache der Beteiligten nicht übereinstimmt: Die Sprache der Entwickler, also der Erfinder, unterscheidet sich in einigen Fällen immens von denen der Patentfachleute.

Zur Verdeutlichung ein Beispiel:

Das Thema der Patentschrift DE19812773(A1) wird von den Entwicklern oder Ingenieuren als Sensor bezeichnet, in der IPC steht jedoch der eher verwirrende Begriff „Halbleiterbauelement und Schwingungserzeuger“. Diese unterschiedliche Bezeichnung ist natürlich für die weitere Arbeitsweise problematisch.

Das heißt, hier musste eine Lösung innerhalb der DaimlerChrysler AG gefunden werden. Zur Lösung kam man durch den so genannten Technologieschlüssel, kurz T-Schlüssel (TS).

---

[31] Patentprofildienst nach der DaimlerChrysler AG (2003):

Der elektronische Profildienst versorgt die Fachbereiche gezielt mit Patentinformationen zu den jeweiligen Fachgebieten. Dazu werden wöchentlich die Datenbanken DE, EP, WO, US nach Ihren individuellen Rechercheprofilen durchsucht.

Dieser wird innerhalb der Patentabteilung angewandt und dient originär der Klassifizierung von Patentfamilien nach bestimmten automobilrelevanten Feldern. Die aktuelle Version beläuft sich hierbei auf ca. 1.500 Einträge, die regelmäßig aktualisiert werden. Der Aufbau der internen Klassifikation wurde hierarchisch realisiert. Die dazu verwendete Baumstruktur ist je nach Gebiet unterschiedlich stark verästelt, die Begriffe jedoch tauchen immer nur an einer Stelle innerhalb der Struktur auf.

Ist kein passender Eintrag vorhanden, wird ein möglichst naher Oberbegriff gewählt. Die Vergabe eines Oberbegriffs und eines darunter liegenden Teilbegriffes ist allerdings nicht zulässig. Beispielsweise darf 1.4.0. und 1.4.1.0. nicht gleichzeitig vergeben werden. Der Technologieschlüssel kann dabei folgendermaßen aufgebaut sein:

1.0.	**Antriebstechnologie (Fahrzeuge)**
1.4.0.	**Verbrennungsantrieb**
1.4.1.0.	Kolbenmotor
1.4.1.1.0.	Gehäuse (Kolbenmotor)
1.4.1.1.1.0.	Zylinderkurbelgehäuse
1.4.1.1.1.1.0.	Zylinder/Zylinderlaufbuchse
1.4.1.1.1.2.0.	Kurbelgehäuse
1.4.1.1.1.3.0.	Lagerdeckel
1.4.1.1.1.4.0.	Kurbelgehäuseunterteil

Abbildung 5: Aufbau des T-Schlüssels am Beispiel der Antriebstechnologie[32]

3.2.3 ePortfolio

Der Begriff „Portfolio“ stammt ursprünglich aus der Börsenwelt. *„Um das Risiko zu verteilen, nutzen die Anleger das Konzept eines Portfolios, also einer ausgewogenen Zusammenstellung einzelner Elemente wie kurzfristige Anleihen, Aktien und langfristige Rentenpapiere, um damit je nach Zielsetzung schnelle, langfristige oder besonders sichere Gewinne zu erzielen.“*[33]

Wie in Kapitel 3.1.2 erläutert, besteht die Gefahr, dass unter Einsatz großer Mittel Innovationen hervorgebracht werden, die sich nachträglich als Doppelerfindungen herausstellen. Daher ist es unablässig, für ein Unternehmen eine ständige Wettbewerbsbeobachtung vorzunehmen. Zusätzlich zu den bisherigen statischen Vorgehensweisen ist deshalb mittlerweile auch eine dynamische Betrachtungsweise möglich. Diese ist beim „ePortfolio-System“, das innerhalb des DaimlerChrysler Konzerns verwendet wird, üblich. Im Rahmen einer Zeitraumbetrachtung können dabei Technik- und Markttrends abgelesen werden. Somit wird ersichtlich, in welchem Bereich ein Defizit besteht und wo weitere Investitionen von Nöten sind.[34]

---

[32] Quelle: Heinz (2004), Folie 23

[33] Wurzer (2003), S. 92

[34] Vgl. Eisenrith (1981), S. 154-157

ePortfolio ist dabei eine flexible Web-Applikation, mit der Endbenutzer schnell auf den Berichtsbestand des Unternehmens zugreifen können.

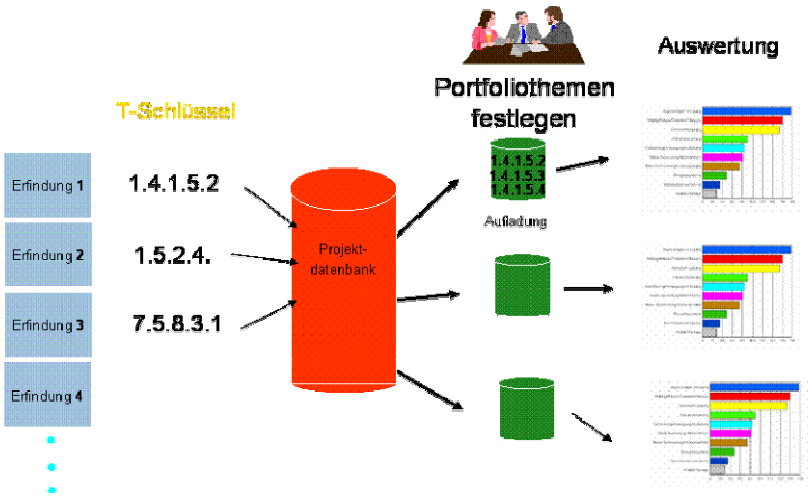


Abbildung 6: Ablauf der Portfoliobildung[35]

Das ePortfolio-Tool greift auf die in Kapitel 3.2.1 genannte Projektdatenbank zu, in der die bereits klassifizierten Patente enthalten sind. Auf Grund der Tatsache, dass in der Projektdatenbank sämtliche Patentschriften, egal ob DaimlerChrysler- oder Fremdschutzrechte, dem dazugehörigen Technologieschlüssel zugewiesen sind, können Statistiken über die Wettbewerber oder auch über das Unternehmensverhalten selber erstellt werden. Dazu wird eine Anfrage der einzelnen Fachbereiche gestellt und *„just in time“*[36] das jeweils für sie wichtige Portfolio erstellt. Die relevanten Daten für die bestellten Themen oder Statistiken werden dann aus der Projektdatenbank in ePortfolio geladen und können anschließend weiterverarbeitet werden.

Die folgenden beiden Abbildungen stellen Beispiele einer erstellten Statistik dar.

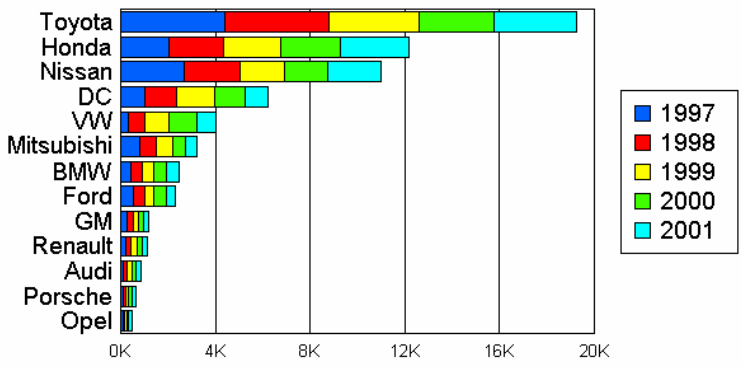


Abbildung 7: Entwicklung der Patent-Erstanmeldungen[37]

---

[35] Quelle: Angelehnt an Heinz (2004), Folie 25

[36] Just in time nach Wikipedia.org (2004):

Just in time ist eine Produktions- und Logistikstrategie. Sie soll Bedarfserfüllungen zum richtigen Zeitpunkt, in der richtigen Qualität und Menge am richtigen Ort gewährleisten.

[37] Quelle: Heinz (2004), Folie 34

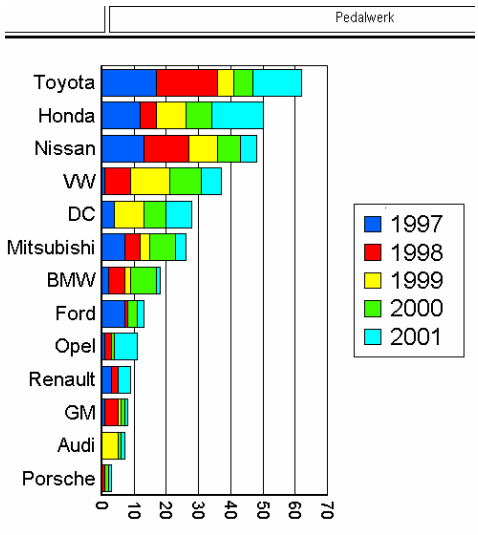


Abbildung 8: Entwicklung der Pedalwerktechnologie-Anmeldungen[38]

3.3 Optimierungspotential

Verbesserungspotenziale des derzeitigen Klassifikationsprozesses und die damit verbundene Problemsituation werden in den folgenden Unterkapiteln aufgezeigt.

3.3.1 Entstehung

Als 2001 die Einführung des Technologieschlüssels beschlossen wurde, wurden in einer groß angelegten Aktion rückwirkend bis 1997 sämtliche eigenen Patentschutzrechte der DaimlerChrysler AG dem Technologieschlüssel zugeordnet. Das erledigte jeder Patent Professional[39] manuell für seine ihm vertrauten Themengebiete. Die Aufgabe bestand darin, jedes einzelne Patentschutzrecht durchzusehen und dem jeweils passenden T-Schlüssel zuzuordnen.

Mit diesen neuen Datensätzen besteht ebenfalls die Möglichkeit, Statistiken zu erstellen, da die Daten auch in ePortfolio geladen werden können.

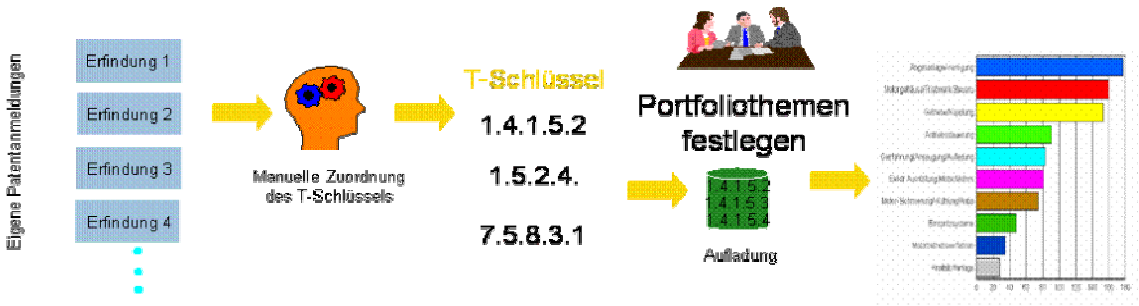


Abbildung 9: Ablauf der manuellen T-Schlüssel Vergabe[40]

[38] Quelle: Heinz (2004), Folie 35

[39] Dazu zählen alle Patentfachleute der Patentabteilung.

[40] Quelle: Heinz (2004), Folie 29

Folgerichtig wurde die Überlegung angestellt, ob die Vorgehensweise der Aufbereitung der internen DaimlerChrysler-Patentschutzrechte nicht auch auf die externen Patente sowie für den fortlaufenden Betrieb angewandt werden könnte. Auf dieser Grundlage sollte ein besserer Vergleich mit den Wettbewerbern möglich sein – vor allem über das ePortfolio-Tool. Eines war dabei allerdings von vornherein klar: Aus Kosten-, Zeit- und Aufwandsgründen konnte und sollte eine manuelle Durchführung nicht erfolgen. Die Klassifikation musste entsprechend automatisch erfolgen.

Dazu wurde ein aufwändiges Programm erstellt, das über verschiedene Schritte das Kalenderwochen-Update-File in die Projektdatenbank schreibt und dabei den T-Schlüssel automatisch den neuen Patentanmeldungen zuordnet. Innerhalb dieses Ablaufs gibt es für die Zuordnung zwei Möglichkeiten, die im Folgenden näher erläutert werden.

3.3.2 Zuordnung der IPC-Klassen zum T-Schlüssel

In einem Vorabschritt wurde in mehreren Durchläufen innerhalb von IPM/C eine Tabelle mit den Übereinstimmungen zwischen Technologieschlüssel und IPC-Klassen erstellt. Ausgangsdaten dazu waren sämtliche Veröffentlichungen von DaimlerChrysler und Toyota der letzten fünf Jahre. Die entstandene Zuordnungstabelle enthält mittlerweile ca. 1.600 Einträge.

Das entwickelte Programm weist nun in einem ersten Durchgang der jeweiligen IPC-Klasse eines Patentschutzrechts den Technologieschlüssel zu.

Zur Verdeutlichung ein Beispiel:

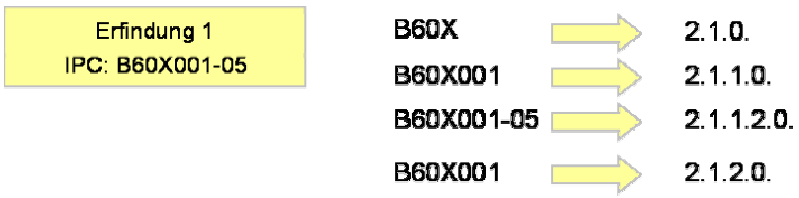


Abbildung 10: Zuweisung einer IPC-Klasse zum T-Schlüssel[41]

Die Erfindung des Wettbewerbers bekommt in dem Fall also 4 T-Schlüssel zugewiesen. Allerdings muss hier noch eine Optimierung erfolgen, indem jeweils die tiefste Ebene ausgewählt wird. Als Endergebnis bleiben somit 2 T-Schlüssel übrig.

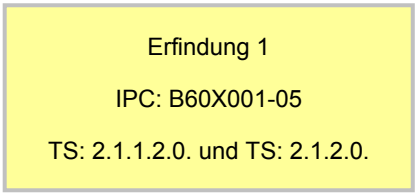


Abbildung 11: Endergebnis für Erfindung 1[42]

[41] Quelle: Angelehnt an Heinz (2004), Folie 30
[42] Quelle: Angelehnt an Heinz (2004), Folie 30

Es treten jedoch Probleme auf, da die Konkordanz[43] bei den mechanischen Technologien sehr gut, bei den nichtmechanischen Technologien jedoch sehr schlecht ist. Z.B. enthält die Klasse B60R016-02 jegliche Themen, die etwas mit Elektrik zu tun haben. Das reicht dabei vom Kabel bis zum Batteriekasten. Auch die Klasse G05B (Steuern und Regeln) weist Probleme auf, da hier sämtliche Patente enthalten sind, die nicht eindeutig zugeordnet werden können. Schwierigkeiten treten auch insofern auf, dass die Klassifikationen durch die Patentämter nicht immer nachvollziehbar sind. Ein weiteres Manko besteht darin, dass die direkte Zuordnung umso schwieriger wird, je tiefer der T-Schlüssel oder die IPC-Ebene verzweigt.

Folglich muss eine Optimierung des Vorgangs erfolgen. Dies geschieht mittels einer Zuweisung des T-Schlüssels zu Rechercheprofilen.

3.3.3 Zuordnung von Rechercheprofilen zum T-Schlüssel

Da die Übereinstimmung mit IPC-Klassen alleine nicht ausreicht, werden die Patentschriften, die nicht über eine IPC-Klasse zuordnungsfähig sind, mit einem Rechercheprofil hinterlegt. Somit werden dem T-Schlüssel bestimmte Schlagwörter zugeordnet. Die Übereinstimmung zwischen T-Schlüssel und Rechercheprofil ist damit genauer, da ein Mitarbeiter die Hinterlegung manuell durchführt. Das heißt, er führt zu einer bestimmten Erfindung eine Recherche in PARS aus und kann alle zu der Erfindung gefundenen Schlagwörter dem T-Schlüssel zuordnen. Alle Treffer des Profils erhalten dann den gleichen Technologieschlüssel. Momentan sind insgesamt 168 Profile zu verschiedenen Technologieschlüsseln hinterlegt.

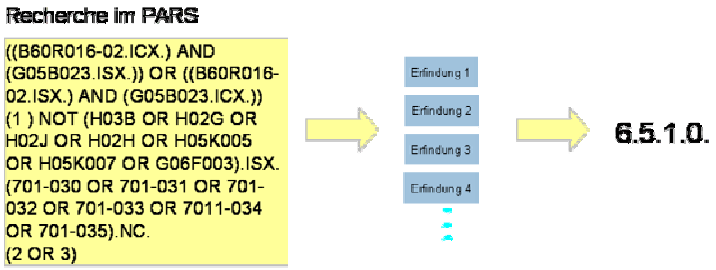


Abbildung 12: Zuweisung eines Rechercheprofils zum T-Schlüssel[44]

Der Prozess weist somit in einem zweiten Durchgang das Rechercheprofil eines Patentschutzrechts dem Technologieschlüssel zu.

Das Optimum an Zuweisungen erfolgt durch Kombination beider Durchgänge

[43] Eindeutige Übereinstimmung der Patentschrift und dem Thema des T-Schlüssels

[44] Quelle: Angelehnt an Heinz (2004), Folie 32

Und an der Stelle tritt das folgende Problem auf:

Ca. 10 Prozent, teilweise auch 20 Prozent der fremden Patentschutzrechte können nach Durchlaufen der beiden Teilprozesse auf Grund einer fehlenden Übereinstimmung keinem Technologieschlüssel zugeordnet werden. In Tabelle 2 sind dazu die erfolgreich klassifizierten und die nicht klassifizierbaren Patentanmeldungen der Wettbewerber von DaimlerChrysler und des Konzerns selber aufgeführt.

Anmelder	Ohne T-Schlüssel	Mit T-Schlüssel
DAIM	2509	26004
BAYM	583	9792
TOYT	4564	60395
VOLS	555	10199
HOND	2609	40234
NSMO	2188	41366
BOSC	3969	49899
DELP	378	4589

Tabelle 2: Erfolgte Klassifizierungen der Patentschutzrechte[45]

Bisher sind die Datensätze so stehen geblieben und wurden nicht weiter beachtet, da

1. keine Lösung bestand
2. die Menge im Gegensatz zu den erfolgreich klassifizierten Daten gering ist.

3.3.4 Manuelle Zuordnung zum T-Schlüssel

Es gibt allerdings eine dritte theoretische Alternative, um die fehlende Klassifikation dennoch zu ermöglichen. Diese besteht aus einer manuellen Zuordnung. Mitarbeiter von IPM/C könnten sich die einzelnen fehlenden Patentschutzrechte ansehen und den Technologieschlüssel ohne Einsatz des oben erwähnten Prozesses zuordnen. Die Übereinstimmung wäre in dem Fall theoretisch am besten, da auf diese Weise sehr genaue Aussagen möglich sind.

Ein derartiger Aufwand wäre allerdings für die Patent Professionals neben ihrer Tagesarbeit nicht zu rechtfertigen. Des Weiteren könnten lediglich um die 60 Dokumente pro Tag geschafft werden. Und der dabei entstehende Kostenfaktor darf in dem Zusammenhang nicht übersehen werden.

[45] Quelle: Heinz (2004), Folie 33

3.3.5 Übereinstimmung der internen DaimlerChrysler-Patentschutzrechte

Ein weiteres Defizit ist bei der Übereinstimmung der Klassifikation der eigenen internen Patentschutzrechte zu finden.

Bei ihrer Anmeldung bekommen die Patentdokumente durch einen Patent Professional mittels einer manuelle Zuordnung den passenden T-Schlüssel zugewiesen, im gleichen Zuge können sie aber auch durch die derzeitige automatische T-Schlüssel-Bestimmung, wie sie für die Fremdschutzrechte erfolgt, einem T-Schlüssel zugewiesen werden. Eine Überprüfung der beiden Alternativen für dieselben Datensätze soll nun zeigen, ob die beiden Zuweisungsmöglichkeiten zum selben Ergebnis führen.

Es stellt sich jedoch heraus, dass es kein fehlerfreies Ergebnis gibt: Auf der höchsten Technologieschlüsselebene ist die Übereinstimmung zwischen beiden Möglichkeiten noch relativ gut, und es wird meist derselbe T-Schlüssel vergeben; je tiefer die Ebenen jedoch untersucht werden, desto größer werden die Differenzen. In Tabelle 3 sind die unterschiedlichen Zuweisungen für die Technologieschlüssel dargestellt. Für jeden T-Schlüssel wird mit der Spalte „Manuell“ angegeben, wie viele Patentdokumente ihm manuell durch einen Patent Professional zugewiesen werden. Mit der Spalte „Zuordnungsprozess“ ist angegeben, wie viele Patentdokumente mittels der automatischen T-Schlüssel-Vergabe zugewiesen werden. Ersichtlich ist daraus, dass man für dieselben Datensätze unterschiedliche Ergebnisse erhält. Z.B. werden für T-Schlüssel 1.0 durch den manuellen Zuordnungsprozess insgesamt 1.835 Patente zugeordnet, mittels des automatischen Zuordnungsprozesses können jedoch nur 1.787 der 1.835 Patente zugewiesen werden. Genau anders herum ist die Situation für Topthema 3. Hier werden durch die automatische Vergabe 720 Patente klassifiziert, mittels der manuellen Vergabe wird jedoch nur eine Teilmenge von 709 Patenten zugeordnet.

T-Schlüssel	Bezeichnung	Zuordnungsprozess	Manuell
1.0.	Topthema 1	1787	1835
2.0.	Topthema 2	3299	3458
3.0.	Topthema 3	720	709
4.0.	Topthema 4	158	187
5.0.	Topthema 5	100	69
6.0.	Topthema 6	406	317
7.0.	Topthema 7	668	618

Tabelle 3: Übereinstimmung der internen Patentschutzrechte auf Ebene 1[46]

[46] Quelle: Eigene Darstellung
Aus Gründen der Geheimhaltung werden die verschiedenen Themen der Patentschutzrechte lediglich unter dem Begriff „Topthema“ angeführt.

Somit ist deutlich erkennbar, dass die Übereinstimmung zwischen dem zweistufigen Zuordnungsprozess und der manuellen Zuordnung durch die Patent Professionals für die höchste Ebene nicht 100 Prozent beträgt. In der nachfolgenden Abbildung wird dieser Zustand mittels eines Diagramms visuell dargestellt.

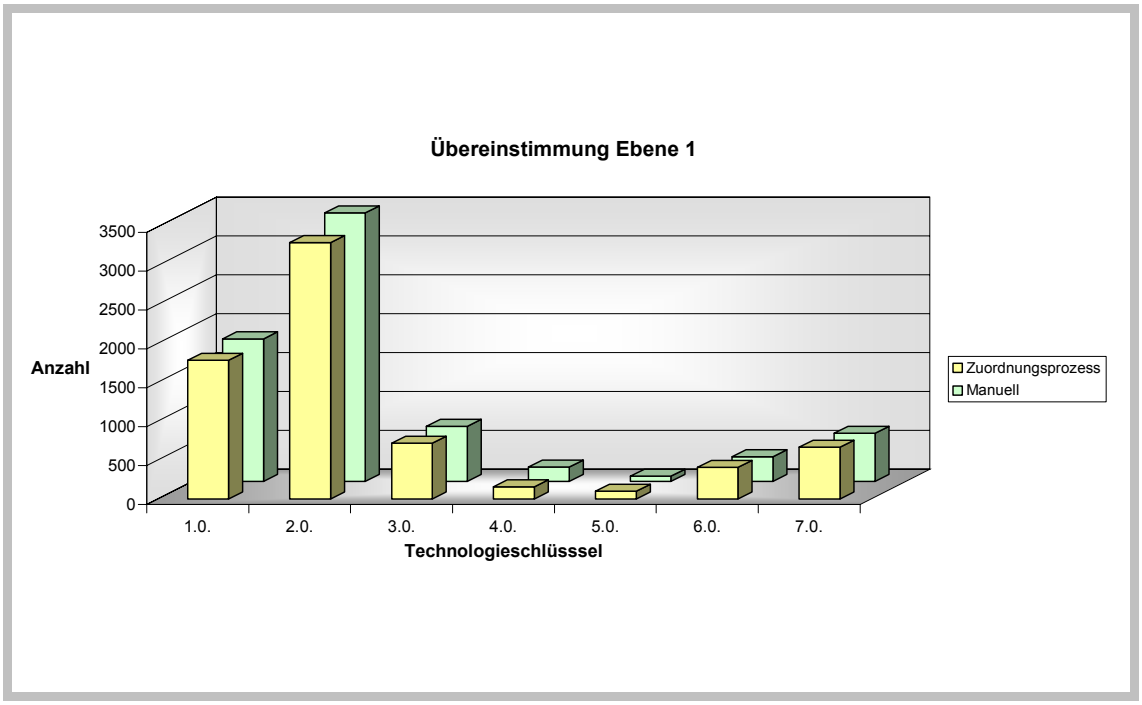


Abbildung 13: Übereinstimmung der internen Patentschutzrechte auf Ebene 1[47]

Bei einer höheren Anzahl von Klassifikationen mittels des Zuordnungsprozesses lässt sich aus Abbildung 13 folgern, dass der Patent Professional bei seiner manuellen Zuordnung nicht ordnungsgemäß vorgegangen ist. Im Gegensatz dazu beweist die höhere Menge der manuellen Zuordnung, dass die 10 bis 20 Prozent der Patente, die nicht durch die zwei Zuordnungsprozesse zugeordnet werden, fehlen.

[47] Quelle: Eigene Darstellung

Die durch den automatischen Zuordnungsprozess ermittelte Menge setzt sich dabei folgendermaßen zusammen:

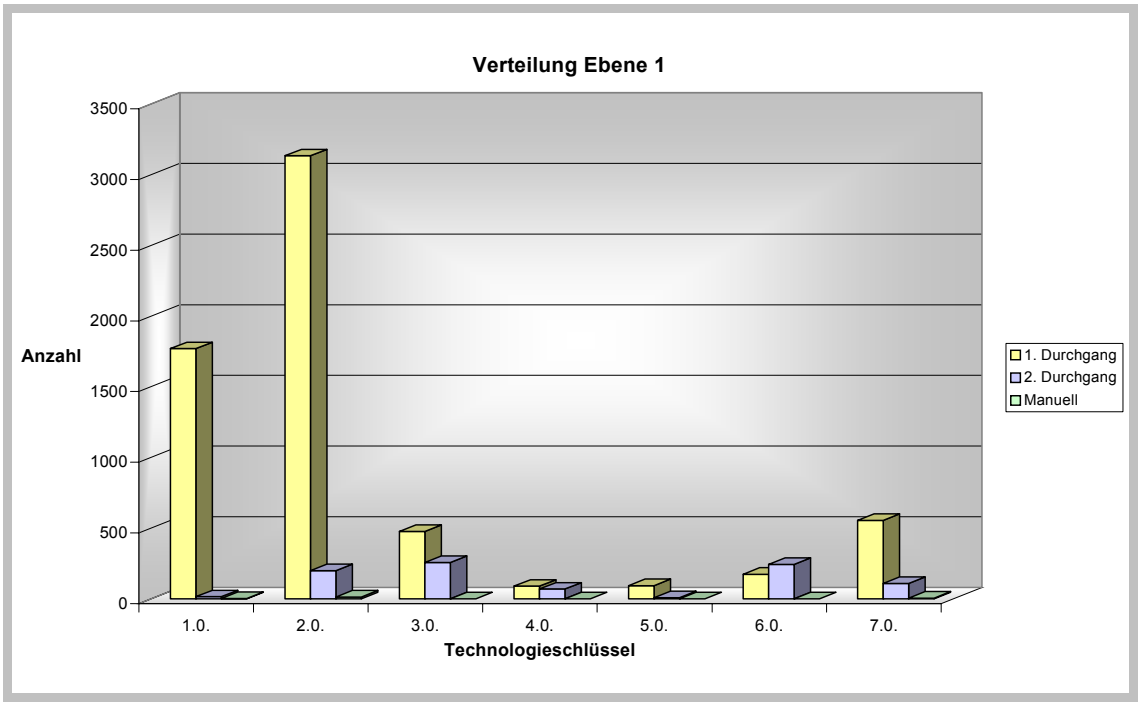


Abbildung 14: Verteilung der internen Patentschutzrechte auf Ebene 1[48]

Deutlich übertrifft die Klassifikation nach Durchgang eins die weiteren Zuordnungen. Darüber hinaus sind Zuordnungen nach Durchgang zwei anzahlmäßig nicht mehr so hoch wie die Übereinstimmung mittels der IPC. Ein minimaler Anteil der zugeordneten Daten kommt auch aus einer manuellen Zuordnung. Das rührt daher, dass für einen kleinen Gesamtteil der Daten aus der Projektdatenbank in früheren Aktionen Datensätze vereinzelt manuell zugewiesen wurden. Diese Zuweisung erfolgte, damit für die anschließende Auswertung durch das ePortfolio-Tool ausschließlich richtige Daten zur Verfügung standen, da es um wichtige strategische Folgeentscheidungen ging.

In der nachfolgenden Tabelle soll die fehlende Übereinstimmung für die zweite Ebene des Technologieschlüssels genauer untersucht werden. Der Unterschied zu der höheren ersten Ebene besteht darin, dass die Datensätze genauer aufgesplittet sind und die Technologieschlüssel weiter verzweigen.

[48] Quelle: Eigene Darstellung

T-Schlüssel	Bezeichnung	Zuordnungsprozess	Manuell
1.1.0.	Topthema 1	2	17
1.3.0.	Topthema 2	27	36
1.4.0.	Topthema 3	936	1122
1.5.0.	Topthema 4	294	360
1.6.0.	Topthema 5	379	255
1.7.0.	Topthema 6	149	38
2.1.0.	Topthema 7	318	323
2.2.0.	Topthema 8	205	11
2.3.0.	Topthema 9	749	753
2.4.0.	Topthema 10	1384	1700
2.5.0.	Topthema 11	84	31
2.6.0.	Topthema 12	36	88
2.7.0.	Topthema 13	432	370
2.8.0.	Topthema 14	75	158
2.9.0.	Topthema 15	16	24
3.1.0.	Topthema 16	544	663
3.2.0.	Topthema 17	134	12
3.3.0.	Topthema 18	1	21
3.4.0.	Topthema 19	41	11
4.1.0.	Topthema 20	158	174
5.1.0.	Topthema 21	13	25
5.2.0.	Topthema 22	14	22
5.3.0.	Topthema 23	64	22
6.1.0.	Topthema 24	128	46
6.2.0.	Topthema 25	20	59
6.3.0.	Topthema 26	20	52
6.4.0.	Topthema 27	210	124
6.5.0.	Topthema 28	28	36
7.1.0.	Topthema 29	387	405
7.3.0.	Topthema 30	124	94
7.4.0.	Topthema 31	6	43
7.5.0.	Topthema 32	138	70
7.6.0.	Topthema 33	3	6

Tabelle 4: Übereinstimmung der internen Patentschutzrechte auf Ebene 2[49]

[49] Quelle: Eigene Darstellung

Deutlich erkennbar ist auch hier, dass die Übereinstimmung nicht mehr so gut wie in Ebene eins ist. Je weiter unten liegende Ebenen betrachtet werden, desto schlechter werden die Ergebnisse.

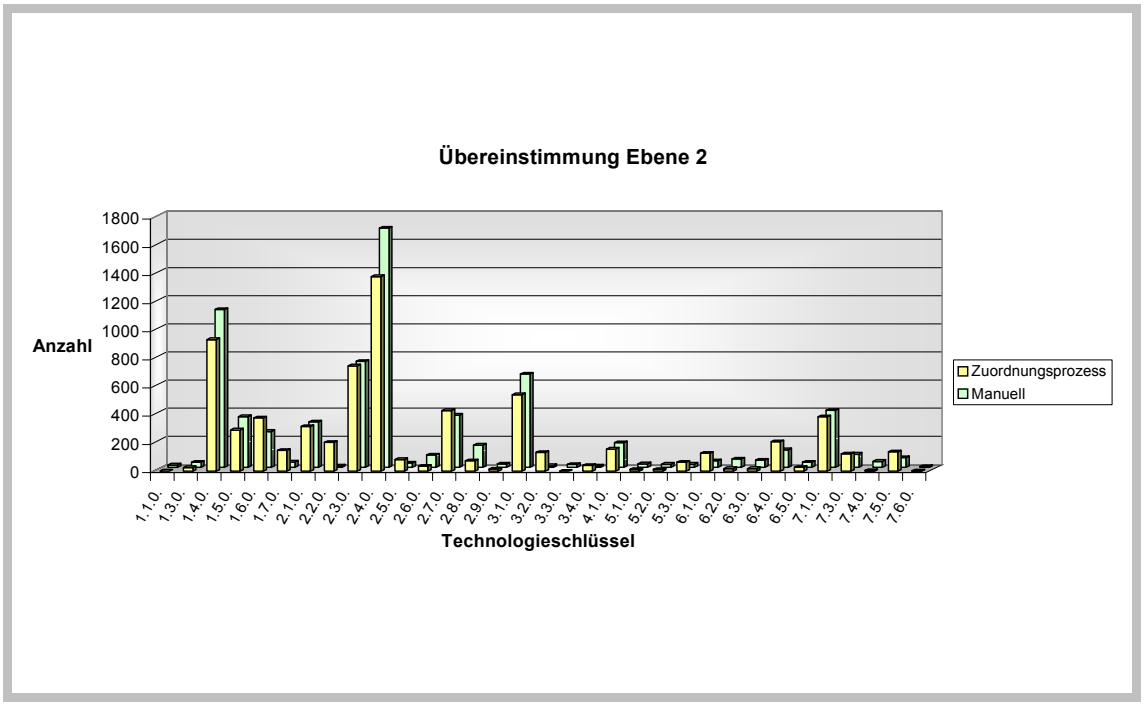


Abbildung 15: Übereinstimmung der internen Patentschutzrechte auf Ebene 2[50]

Die durch den automatischen Zuordnungsprozess ermittelte Menge setzt sich dabei folgendermaßen zusammen:

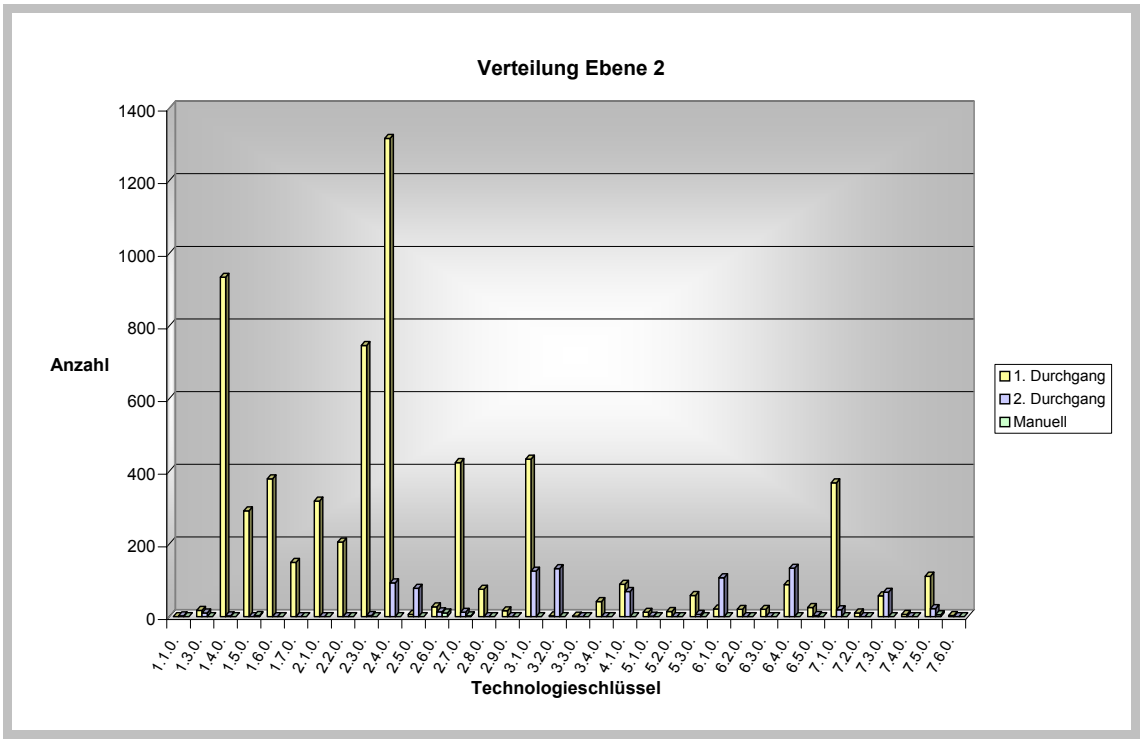


Abbildung 16: Verteilung der internen Patentschutzrechte auf Ebene 2[51]

[50] Quelle: Eigene Darstellung

Erneut ist die Zuordnung durch den 1. Zuordnungsdurchgang deutlich am höchsten. Die Menge der Patente, die durch ein Rechercheprofil zugeordnet sind, ist teilweise jedoch höher als bei Ebene eins. Abermals minimal ist die manuelle Zuordnung.

3.4 Fazit

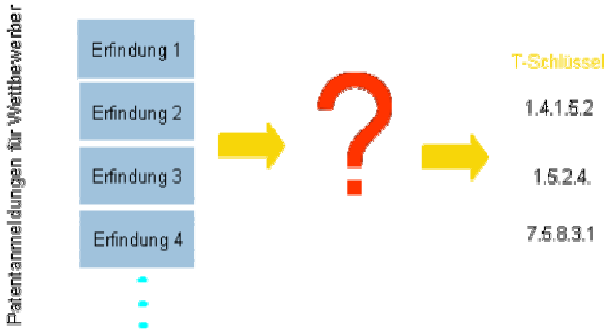


Abbildung 17: Gesucht wird nach einer Lösung[52]

Letzten Endes sucht IPM/C eine Lösung bzw. eine neue Methode, die besser als der derzeitige Realisierungsprozess ist und ihn von der Qualität her übertrifft. Dabei soll vor allem der 10%-ige Fehleranteil beseitigt werden.

Als möglicher Lösungsweg wird ein Text-Mining-Verfahren in Betracht gezogen, das die Patentanmeldungen automatisch dem Technologieschlüssel zuweisen soll.

3.5 Benutzerkreis

Die Idee geht dahin, dass zwei Benutzergruppen mit dem System arbeiten sollen:

a.) Die Administratoren des Systems. Das sind die Mitarbeiter von IPM/C, die die Daten vorbereiten und die Klassifikation überprüfen bzw. die Systemadministration übernehmen.

b.) Mitarbeiter von DaimlerChrysler, die sich einen kurzen und schnellen Überblick über bestehende Patentschutzrechte machen wollen, ohne große Fachkenntnisse in einer Recherchesprache haben zu müssen, denn dafür kann das bestehende Patentinformationssystem PARS genutzt werden. Somit können auch ungeübte Benutzer und Laien ihre Informationsbedürfnisse befriedigen.

Aus dem sich gezeigten Verbesserungspotenzial ergeben sich im Folgenden die spezifischen Anforderungen der Patentabteilung an ein Text-Mining-System. Des Weiteren wird ein Überblick über allgemeingültige Anforderungen gegeben.

---

[51] Quelle: Eigene Darstellung

[52] Quelle: Angelehnt an Heinz (2004), Folie 28

3.6 Anforderungen

„Unter einer Anforderung wird [dabei] eine Aussage über eine zu erfüllende Eigenschaft oder zu erbringende Leistung eines Produkts verstanden, über einen Prozess oder der am Prozess beteiligten Personen.“[53]

3.6.1 Anforderungen von IPM/C

Gefordert wird ein qualitativ und quantitativ hochwertiges System das folgende funktionale Anforderungen erfüllen soll:

1. Erleichterung

Der Klassifikationsprozess soll vereinfacht und somit die Probleme der beiden Zuordnungsdurchgänge behoben werden. Dazu zählt unter anderem die aufwändige Pflege der Rechercheprofile. Ferner ist gewünscht, dass die theoretisch erwogene manuelle Zuordnung nicht mehr in Betracht gezogen werden muss und das System automatisch alle Patentschutzrechte einem Technologieschlüssel zuweist.

2. Wirtschaftlichkeit

Letztendlich tritt dadurch der Faktor der Wirtschaftlichkeit auf: Kosten und Zeit sollen durch die erleichterte Pflege eingespart werden. Und auch auf Grund der Tatsache, dass im Endeffekt alle internen und externen Patentanmeldungen vollständig in der Projektdatenbank enthalten sind, soll eine optimale Nutzung des Statistik-Tools ePortfolio gewährleistet werden.

3. Fehlerbeseitigung

Als weitere Anforderung soll die bestehende Fehlerrate von 10 bis 20 Prozent verbessert bzw. im Idealfall gänzlich beseitigt werden. Alle relevanten Patente sollen abgefangen und dementsprechend zuverlässig klassifiziert werden können.

4. Beschleunigte Informationsaufnahme

Das System soll darüber hinaus auch als Informationsquelle genutzt werden. Das Tool wirkt somit komprimierend, da eine größere Datenmenge zu spezifischem Wissen verdichtet werden kann.

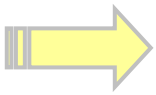
Auf Grund der Möglichkeit, Informationen in Dokumenten zu organisieren und zu strukturieren, soll es dem Benutzer möglich sein, sich schneller mit einem Thema vertraut zu machen. Da Verknüpfungen zwischen identifizierten Objekten erkannt werden, wird auch das „Wissen zwischen den Zeilen erkannt“, welches oft wesentlich zum Verständnis eines Sachverhaltes beiträgt. Die Recherche wird auf diese Weise vereinfacht, da der Nutzer keine schwierige Syntax des Recherchesystems beherrschen muss.

---

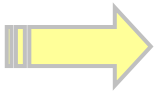
[53] Rupp (2001), S. 10

Ein typisches Beispiel dazu wäre, dass ein Anwender bezüglich einer Patentanmeldung alle dazu bereits existierenden Patente schnell auf einen Blick sehen oder alle ähnlichen Patente der Mitbewerber herausfinden möchte.

Zusammenfassend sollen vor allem zwei Dinge mit dem System möglich sein:



Die Verbesserung der Effizienz[54] der Patentklassifikation



Die Verbesserung der Effektivität[55] der Recherche

3.6.2 Allgemeingültige Anforderungen

Bei einer Evaluierung wird im Normalfall nicht ausschließlich die Funktionalität geprüft, wie es in dem vorliegenden Projekt der Fall ist, sondern eine umfassende Untersuchung der Software-Ergonomie[56] durchgeführt. Der Grund für dieses eine überprüfte Kriterium liegt darin, dass mit dem vorliegenden Testlauf für IPM/C geklärt werden soll, ob eine umfangreiche Investition in das Text-Mining-Tool überhaupt sinnvoll ist. Im Falle eines positiven Ergebnisses wird in einer sich anschließenden Evaluierungsphase auf weitere Kriterien geachtet, die nach Angaben der Forschungseinheit RIC/AM umgesetzt werden könnten. Norbert Gronau[57] schlägt für die Gliederung einer solchen Anforderungsspezifikation mögliche Einteilungen vor:

- Technische Anforderungen

Zu den technischen Anforderungen zählt unter anderem die Eingliederung in bestehende Systemlandschaften oder die Schnittstellenbewältigung im Hinblick auf andere im Unternehmen eingesetzte Applikationen.

---

[54] Effizienz nach net-Lexikon.de (2004):
Effizienz ist das Verhältnis eines in definierter Qualität vorgegebenen Nutzens zu dem eingesetzten Aufwand, der zur Erreichung des Nutzens nötig war. Als effizientes Verhalten bezeichnet man ein Verhalten, das nicht nur zur Erreichung eines gegebenen Zieles führt, sondern dabei den Aufwand gering hält.

[55] Effektivität nach net-Lexikon.de (2004):
Effektivität ist das Verhältnis aus definiertem Ziel und erreichten Ziel. Ein Verhalten ist dann effektiv, wenn es ein vorgegebenes Ziel erreicht, es ist wenig effektiv, wenn das Ziel nicht oder nur teilweise erreicht wird.

[56] Software-Ergonomie nach Herczeg (1994), S. 254:
Es handelt sich dabei um eine interdisziplinäre Wissenschaft mit dem Ziel der benutzer- und aufgabengerechten Gestaltung von Arbeit mit interaktiven Computersystemen.

[57] Gronau (2001), S. 113-115

❖ Anforderungen an die Benutzerfreundlichkeit (Usability)[58]

Diese Anforderung soll sicherstellen, dass die Software *„mit einem Minimum an Schulungsaufwand“*[59] von den Mitarbeitern bedient werden kann. Hinsichtlich der Benutzerfreundlichkeit wird auf ein übersichtliches und leicht zu bedienendes System Wert gelegt. Da auch für die Administration so wenig Aufwand wie möglich betrieben werden soll, wird ein System mit möglichst wenig Administrationsaufwand bevorzugt.

❖ Adaptive Anforderungen

Unter adaptiv versteht Norbert Gronau[60] schließlich *„die Fähigkeit […] sich an veränderte organisatorische Bedingungen anzupassen*.“ Dazu zählt z.B. die Flexibilität: Danach sollte ein System nicht nur die bisherigen Anforderungen des Unternehmens berücksichtigen und unterstützen. Vielmehr muss es in der Lage sein, sich an die stetig veränderten Bedürfnisse des Marktes anzupassen. In dem vorliegenden Projekt von IPM/C wäre das z.B. eine Anpassung an die sich kontinuierlich verändernde IPC-Klassifikation.

Die Zusammenarbeit des Benutzers mit dem Computersystem wird dabei durch die Dialogschnittstelle unterstützt, die sich nach ISO 9241[61] an folgenden Gestaltungsgrundsätzen orientieren sollte:

❖ Erwartungskonformität

Hier wird darauf geachtet, ob das System bei den Antworten mit längeren Bearbeitungszeiten reagiert oder ob das System die Orientierung durch eine uneinheitliche Gestaltung erschwert oder es sich nicht durchgehend nach einem einheitlichen Prinzip bedienen lässt.

❖ Aufgabenangemessenheit

Dabei handelt es sich um die Unterstützung des Benutzers zur effektiven und effizienten Erledigung seiner Aufgabe. D.h., ein System ist dann aufgabenangemessen, wenn es den Nutzer bei der Erledigung seiner Aufgaben unterstützt, ohne ihn unnötig zu belasten.

❖ Selbstbeschreibungsfähigkeit

Darunter wird die Möglichkeit des Benutzers verstanden, einzelne Dialogschritte nachvollziehen zu können, entweder durch Rückmeldung des Systems oder durch konkrete Nachfrage. D.h., das System ist dann selbstbeschreibungsfähig, wenn es dem Nutzer auf Anforderung Erläuterungen zu jedem einzelnen Dialogschritt liefern kann.

---

[58] Usability nach ISO 9241 (1997):

Die Usability eines Produkts ist dabei das Ausmaß, in dem es von einem bestimmten Benutzer verwendet werden kann, um bestimmte Ziele in einem bestimmten Kontext effektiv, effizient und zufriedenstellend zu erreichen.

[59] Gronau (2001), S. 113

[60] Gronau (2001), S. 115

[61] Scoreberlin GmbH (1999-2004)

- Steuerbarkeit

Darunter wird verstanden, dass der Benutzer in der Lage ist, den Dialogablauf zu starten sowie dessen Richtung und Geschwindigkeit bis zur Zielerreichung zu beeinflussen. Der Nutzer kann also in die Systemabläufe aktiv eingreifen, z.B. indem er Aktionen rückgängig macht. Der Nutzer soll in der Lage sein, die Dialoge kontrollieren zu können.

- Erwartungskonformität

Der Dialog ist konsistent und entspricht den Merkmalen des Benutzers, die gekennzeichnet sind durch die Kenntnisse im Arbeitsgebiet, der individuellen Ausbildung und Erfahrung sowie durch allgemein anerkannte Normen. Es ist also erwartungskonform, wenn es den Erwartungen des Nutzers entspricht.

- Fehlertoleranz

Ein System ist dann fehlerrobust, wenn der Nutzer trotz fehlerhafter Bedienung das beabsichtigte Arbeitsergebnis trotzdem nur mit geringem Korrekturaufwand erreicht. Dazu muss ihm der Fehler zunächst in verständlicher Form mitgeteilt bzw. - soweit wie möglich - automatisch durch das System selbst behoben werden. Das Arbeitsergebnis kann trotz fehlerhafter Eingabe seitens des Benutzers ohne großen Mehraufwand erreicht werden.

- Individualisierbarkeit

Darunter wird die Anpassungsfähigkeit des Systems an den Benutzer verstanden.

- Lernförderlichkeit

Unterstützung des Benutzers beim Erlernen des Systems, beispielsweise durch eine entsprechende Hilfefunktion.

4 Stand der Technik

In diesem Kapitel erfolgt die Vorstellung des Text-Mining und dessen thematische Einordnung in das Gebiet der Informationsgewinnung. Das Konzept der Arbeitsweise von Text-Mining-Systemen wird im Überblick vorgestellt.

Nach Jochen Dörre et al.[62] ist das Konzept der Volltextsuche (engl. Information Retrieval) und des Text-Mining in den letzten Jahren auf Grund der zunehmenden Bedeutung des Intranet und Internet in Organisationen sehr populär geworden. Das hat seine Ursache in der Zunahme der für einen Benutzer verfügbaren Datenflut.

4.1 Information Retrieval

Ausgangspunkt und Basis des Text-Mining ist das Dokumentenmanagement und hier insbesondere der Bereich des Information Retrieval.

Information Retrieval wurde historisch gesehen zum besseren (Wieder)auffinden von wissenschaftlicher Literatur entwickelt. „*Retrieval ist somit die Methode, in einem bestimmten Datenbestand Suchvorgänge durchzuführen*."[63]

Aber nicht nur das „wiederauffindbar machen" zählt zu dem Begriff, sondern auch der Zugriff, die Art der Darstellung, die Speicherung und die Organisation der Informationen gehören dazu. So lässt sich der Prozess des Information Retrieval zweiteilen: Auf der einen Seite muss ein Index konstruiert werden, der alle Dokumente umfasst, die Ziel der Suche sein können, und das eigentliche Retrieval auf der anderen Seite.
Das verwendete Prinzip ist dabei relativ einfach gehalten: Bei einer Analyse sämtlicher Texte, die gesucht werden können, werden alle im Text vorkommenden relevanten Fachwörter samt ihren Positionen ermittelt und in einer geeigneten Datenstruktur, Index genannt, als Indexterme gespeichert. Vergleichbar ist diese Methode mit dem Schlagwortverzeichnis am Ende eines Buches. Eine Anfrage an das System entspricht somit dem Nachschlagen im Index.[64]

Dokumente werden demnach durch Bezeichnungen in der Dokumentenkollektion repräsentiert. Der Nutzer formuliert anschließend eine Suchanfrage, die mit den Bezeichnungen der Dokumente abgeglichen wird. Stimmt die Bezeichnung eines Dokuments mit der gestellten Suchanfrage überein, so wird das für die Suchanfrage relevante Dokument in einer Ergebnismenge dem Nutzer präsentiert.[65]

---

[62] Dörre et al. (2001), S. 425
[63] Poetzsch (2001), S. 13
[64] Vgl. Dörre et al. (2001), S. 425
[65] Vgl. Haag (2002), S. 16

Abbildung 18 stellt nochmals einen Überblick der Komponenten und des Datenflusses einer typischen Text-Retrieval-Architektur dar:

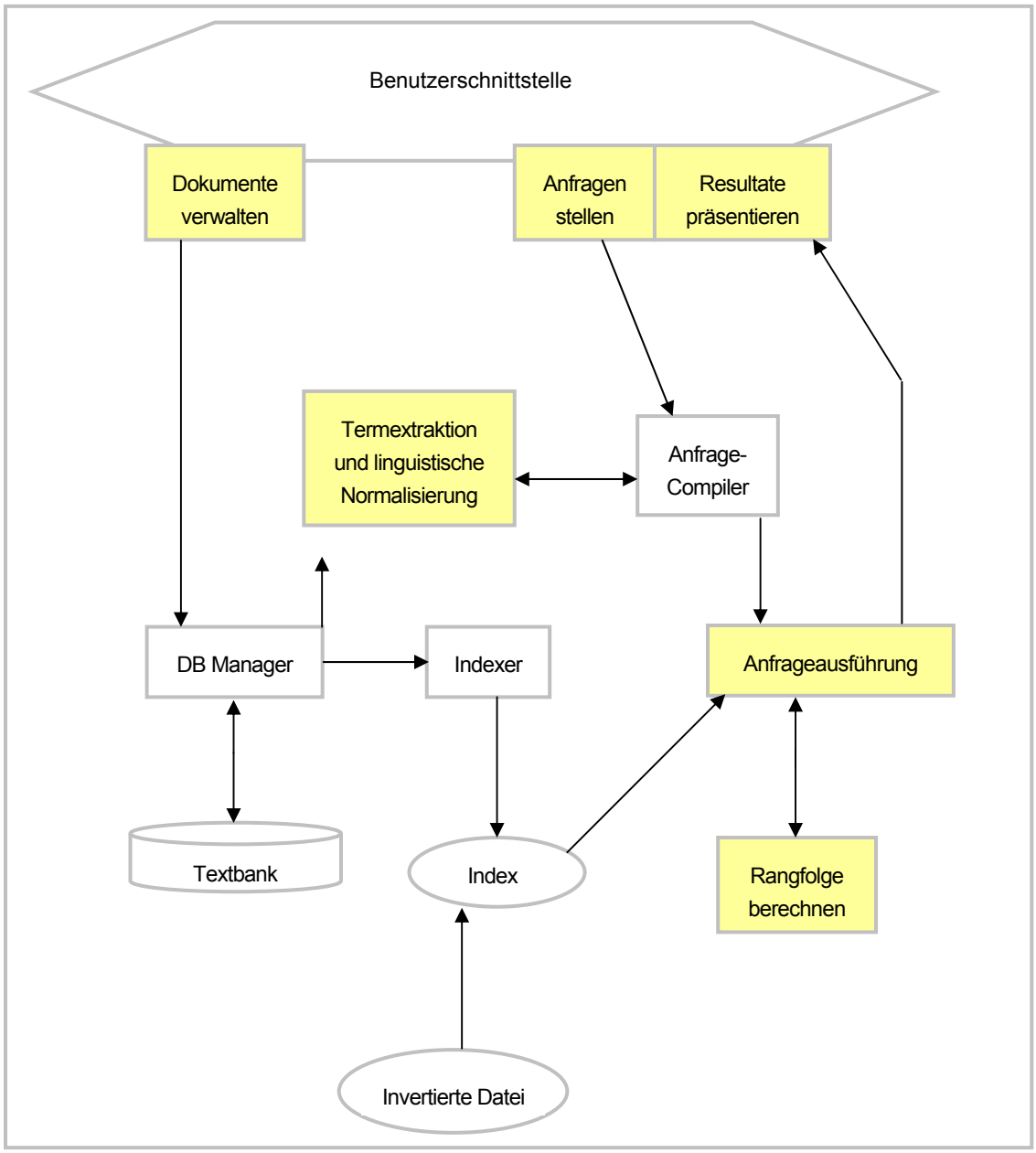


Abbildung 18: Komponenten eines Volltextsuchsystems[66]

Zu den Merkmalen eines Information Retrieval Systems zählen dabei:

- Die Summe ist sehr groß - im Normalfall handelt es sich um Tausende/Millionen Seiten von Text.
- Die Summe ist relativ statisch zur Anfragehäufigkeit.
- Eine Anfrage muss in sehr kurzer Zeit beantwortet sein - im Idealfall im Sekundenbereich.

[66] Quelle: Dörre et al. (2001), S. 429

Deshalb kommt es auf der Seite der Sortierung und Speicherung der Indexterme und ihrer Positionen darauf an, mit großen Datenbeständen effizient und sicher umgehen zu können. Das wird mit Verfahren aus dem Arsenal der Datenbanktechnologie bewerkstelligt. Auf der Seite der Retrievals hingegen zählt die formale Semantik der Anfrage. Durch so genannte Retrievalmodelle wird dabei festgelegt, wie Terme und Operatoren in einer Anfrage zu interpretieren sind. Zu den bedeutendsten Modellen zählen das Boolesche Retrievalmodell und das Vektormodell. Beim Booleschen Modell bestehen die Anfragen aus Termen, die mit Booleschen Operatoren wie *UND*, *ODER* und *NICHT* verknüpft werden können. Einen völlig anderen Ansatz als das Boolesche Modell wählt das Vektormodell oder Vektorraummodell. Es bestimmt Ähnlichkeiten zwischen Dokumenten oder zwischen Dokumenten und einer Anfrage als Abstand oder Winkel zwischen deren Vektoren. Auf das Verfahren wird näher in Kapitel 4.4.3.5.2 eingegangen.

4.2 Data Mining

Nach Sascha Lorenz[67] wird unter Data Mining, auch „Knowledge Discovery in Databases“ bezeichnet, die Extraktion implizit vorhandenen, nichttrivialen und nützlichen Wissens aus großen, dynamischen und relativ komplex strukturierten Datenbeständen oder Datenbanken verstanden. Insbesondere wirtschaftliche Kennzahlen werden dazugerechnet. Ein in der Praxis häufig anzutreffendes Beispiel findet sich im Marketing, wo der Umsatz bestimmter Artikel mit den jeweiligen Gruppen von Käufern in Verbindung gebracht wird.

Die bekanntesten Verfahren sind das so genannte Clustering zur Segmentierung der Daten nach ihren ähnlichen Eigenschaften und die Klassifikation zur Analyse bzw. Vorhersage von Werten für einzelne Datenelemente.[68]

Data Mining und Text-Mining teilen sich die eingesetzten Verfahren und Algorithmen[69]. Data-Mining-Algorithmen können jedoch nicht direkt auf das Text-Mining angewendet werden.

---

[67] Lorenz (2001), S. 8

[68] Vgl. Dörre et al. (2001), S. 425

[69] Algorithmus nach net-Lexikon.de (2004):

Ein Algorithmus ist eine Verfahrens- oder Verarbeitungsvorschrift; zum Beispiel für einen Rechenvorgang, der wiederholt nach einem bestimmten, gleich bleibenden Schema abläuft. Die Verarbeitungsvorschrift muss dabei so präzise formuliert sein, dass sie von einer Maschine (etwa einem Computer) durchgeführt werden kann. Ein Algorithmus muss also eindeutig sein und ein klar definiertes Ende haben.

4.3 Informationsextraktion

Je mehr Text online zur Verfügung steht, desto schwieriger wird es, das Informationspotenzial gezielt zu nutzen, d.h., relevante Informationen zu finden, zu extrahieren und in kompakter Form darzustellen.

Eine sich dabei neu etablierte Forschungsrichtung ist die Erforschung und Realisierung von Systemen zur Informationsextraktion. Das Ziel dabei ist die Konstruktion von Verfahren, die gezielt Informationen aus freien Texten aufspüren und strukturieren, bei gleichzeitigem Überlesen von irrelevanten Informationen. Das Verfahren versucht keine umfassende Analyse des gesamten Inhalts aller Textdokumente, sondern es will nur die Textpassagen analysieren bzw. „verstehen“, die relevante Informationen beinhalten. Auf diese Weise ist die Identifikation und Markierung von Einheiten eines Textdokuments möglich, die für die weitere Informations- und Wissensaneignung von besonderer Bedeutung sind.

Die so extrahierten Daten können vielseitig eingesetzt werden - z.B. zur Textfilterung oder Textklassifikation, als Einträge für Datenbanken, zur Unterstützung von Text-Mining-Systemen oder als Ausgangspunkt für eine Textzusammenfassung.[70]

4.4 Text-Mining

Das Ziel des Text-Mining ist die Erfassung des Inhalts von Dokumenten. Dabei soll unter dem Begriff „Dokument“ nicht nur ein bestimmter Text verstanden werden, sondern eine gesamte Textmenge. Deshalb ist auch ein Abschnitt eines Textes ein Dokument. Mittels Text-Mining können somit wertvolle Informationen in großen Mengen von Daten gesucht werden.

4.4.1 Einordnung des Text-Mining

In der Literatur sind dazu jedoch weder einheitliche Prozessmodelle noch eine eindeutige Abgrenzung zwischen den Methoden des Text-Mining und den Disziplinen Information Retrieval, Data Mining und Informationsextraktion beschrieben.

Nach Günter Neumann[71] ist Text-Mining aber ein verwandtes Forschungsgebiet zur Informationsextraktion, da ähnliche Ziele angestrebt werden - dennoch unterscheiden sich die verwendeten Methoden erheblich. Darüber hinaus wendet Text-Mining auch die bereits bei strukturierten Datenmengen erfolgreich benutzten Techniken des Data Mining an. Im Unterschied zum Data Mining arbeitet das Text-Mining jedoch nicht auf den Daten einer Datenbank, sondern auf Textdokumenten. Die Analyseaufgaben und Analyseziele sind aber ähnlich.

---

[70] Vgl. Krause (2002), S. 584

[71] Neumann (2001), S. 448

Das Spektrum der vorhandenen Methoden und Systeme ist sehr breit und vielfältig. Es reicht von sehr gut erforschten regelbasierten oder wissensbasierten Techniken bis hin zu statistischen Systemen. Daneben gibt es noch die Einteilung in lernende und nicht-lernende Verfahren, wobei die statistischen Verfahren in der Regel lernfähig sind.

Regelbasierte Verfahren haben gegenüber den statistischen einen Vorteil: Sie sind für den Nutzer transparent. Außerdem können neue Regeln relativ leicht erstellt oder angepasst werden. Ein Nachteil liegt allerdings darin, dass die verwendeten Verfahren von Hand erstellt werden. Sie können somit nicht lernen. Ihre Intelligenz steckt in den Regeln, die ein Mensch definiert hat. Ein weiteres Manko: Regelbasierte Algorithmen ordnen Dokumente einer Kategorie zu oder nicht. Statistische Verfahren liefern hingegen Wahrscheinlichkeiten zwischen null und eins. Das ist von Vorteil, weil man damit den so genannten Schwellwert[72] einstellen kann, ab dem ein Dokument einer Klasse angehören soll oder nicht.

Thomas Kamphusmann[73] erklärt, dass Text-Mining somit die Möglichkeit bietet, schnellere, flexiblere und präzisere Recherchen durchzuführen. Ferner unterstützt es die Organisation und Pflege der Textbestände. Zudem kann ohne präzises Vorwissen hinsichtlich der Inhalte und der Strukturen eine Recherche in großen Archiven durchgeführt werden. Hier steht nicht die gezielte kontextabhängige Suche und noch weniger die Kategorisierung großer Mengen im Vordergrund, sondern die Ermöglichung eines surfenden Zugangs, der je nach den ersten Ergebnissen weitere Recherchen nach sich zieht. Idealerweise sollte der gesamte Weg - angefangen bei einer nur vagen Vorstellung des Gesuchten bis zum Fund des passenden Dokuments - eine sich laufend anpassende Präzisierung der Suche sein.

Anders gelagert sind Szenarien, in denen regelmäßig große Mengen von Dokumenten aufbereitet werden müssen. Hierbei steht weniger die Unterstützung zielgerichteter Suchen im Vordergrund, sondern die Entlastung von regelmäßigen Arbeiten wie die Sortierung von eingehenden eMails, Nachrichten etc.

Der Vorteil des Text-Mining gegenüber herkömmlichen Suchverfahren ist immens: Gibt die suchende Person etwa als Stichwort Begriffe aus ihrem Text wie „Motor“, „Optimierung“ und „Benzineinsparung“ ein, so würde sie mit Sicherheit einen Text nicht finden, der dieselbe Thematik behandelt, diese aber mit anderen Stichwörtern wie „Antrieb“ und „Verbrauchsreduktion“ charakterisiert hat. Dem Benutzer würden nicht nur relevante Dokumente verborgen bleiben, die Suchmaschine würde ihm umgekehrt auch viele Dokumente anbieten, die für seine aktuelle Recherche unwichtig sind. Beide Schwächen solcher Suchmaschinen minimiert dagegen das Text-Mining.[74]

---

[72] Z.B. kann bestimmt werden, dass bei einer Ähnlichkeit von 15 Prozent alle Treffer unterhalb des Werts abgeschnitten werden und somit in der Ergebnisliste nicht auftauchen.

[73] Kamphusmann (2002), S. 39-41

[74] Vgl. DaimlerChrysler AG (2002), S. 37

4.4.2 Aufgaben des Text-Mining

Verfahren des Text-Mining helfen, Suchergebnisse zu analysieren und geben Hinweise zur Verbesserung der Suchanfrage. Darüber hinaus lassen sie sich optimal einsetzen, um neue Zusammenhänge aufzudecken und für strategische Zwecke aufzubereiten. [75] Im Folgenden werden einige Teilaspekte des Text-Mining erläutert, die als Grundlage für eine Fülle von weiteren Verfahren dienen, von denen im Rahmen der Arbeit nur ein kleiner Ausschnitt beleuchtet werden kann.

Das Hauptziel ist, weitgehend automatisiert aus großen Dokumentenkollektionen aussagekräftige Muster sowie Inhalte zu identifizieren und sie dem Anwender komprimiert als interessantes Wissen zu präsentieren. Text-Mining macht das Lesen von Texten aber nicht völlig überflüssig, sondern leitet den Anwender zu potenziell interessanten Aussagen. Somit liegt der Nutzen darin, dass die in einer umfangreichen Dokumentensammlung verborgenen Fakten in Entscheidungssituationen mit geringem Zeitaufwand erschlossen werden können.[76]

4.4.2.1 Textsuche

Das gebräuchlichste Einsatzgebiet von Text-Mining ist die Suche nach Texten innerhalb einer größeren Sammlung gemäß der Anfrage eines Benutzers. Wenn diese Anfrage aus einzelnen Wörtern oder Sätzen besteht, handelt es sich um eine klassische Information-Retrieval-Funktion. Mit dieser Funktion ist jeder Internetnutzer vertraut, z.B. in Form von Suchmaschinen, die dazu verwendet werden, interessante Dokumente zu finden. Wenn die Suche nicht nur auf Schlagwörtern basiert, sondern mit Beispieldokumenten durchgeführt wird, entspricht die Funktion schon eher dem Text-Mining. In dem Fall wird eine gesamte Kollektion analysiert, um daraus einen oder mehrere Texte auszugeben, die dem Interessengebiet des Benutzers entsprechen.

4.4.2.2 Informationsextraktion

Eine weitere bemerkenswerte Text-Mining-Funktion ist die Extraktion von Informationen. Bei einer normalen Suchanfrage ist es üblich, dass hunderte von Treffern gefunden und dem Benutzer ausgegeben werden. Um nun aus dieser Anzahl von gefundenen Texten entscheiden zu können, ob ein Text überhaupt relevant ist, ist eine kurze und aussagekräftige Beschreibung des Textes von Nöten. Diese charakteristische Beschreibung kann in Form von wenigen Schlagwörtern, aus mehreren textähnlichen Begriffen oder aus einer Zusammenfassung des Textes vorliegen. Das Ziel besteht darin, nicht die gesamte Kollektion zu analysieren, sondern nur relevante Informationen über den Text auszugeben.

---

[75] Vgl. Dörre et al. (2001), S. 434

[76] Vgl. Meier/Beckh (2000), S. 165

4.4.2.3 Analyse von Textkollektionen

Die dritte Gruppe von Text-Mining-Aufgaben zielt darauf ab, eine Übersicht über eine Textkollektion anzubieten. Dazu gehören unter anderem Klassifikation und Clustering. Entschieden werden soll durch die Analyse, zu welcher vorgegeben Kategorie ein Text gehört. Die Analyseaufgabe beim Clustering besteht hingegen darin, dass Texte ohne jegliche Kategorienzugehörigkeit zur Textgruppierung verwendet werden.[77]

4.4.3 Methoden des Text-Mining

Zu den Methoden des Text-Mining wird eine Vielzahl an Techniken gezählt, auf die im Einzelnen nicht ausführlich eingegangen werden kann. Jedoch soll eine exemplarische Vorstellung der Verfahren erfolgen, die im Zusammenhang mit der Arbeit und der weiteren Evaluierung von Bedeutung sind.

4.4.3.1 Stemming oder Lemmatisierung

In natürlichsprachigen Texten kommen zahlreiche Varianten eines Wortes vor. Da die einzelnen Varianten letztendlich den gleichen Begriff beschreiben, ist es für Text-Mining-Anwendungen zur Steigerung der Effektivität wichtig, möglichst alle Varianten eines Wortes zu ermitteln. Verschiedene Techniken zur Durchführung dieser Verschmelzung werden unter dem Begriff des „Stemming“ oder der „Lemmatisierung“ zusammengefasst.[78]

Beim Stemming oder der Stammformreduktion werden die Wortformen auf ihren Stamm zurückgeführt. Diese Form ist im Allgemeinen keine in der Sprache als Wort vorkommende Form und kann z.B. für ein Verb und ein Substantiv gleich sein.[79]

4.4.3.2 Thesaurus

Laut Reginald Ferber[80] erfassen Thesauren Wörter, Terme und Ausdrücke eines Sachgebiets und beschreiben die Beziehungen zwischen ihnen. Thesauren haben vor allem zwei Funktionen:

a.) Sie definieren ein kontrolliertes Vokabular und

b.) sie stellen Beziehungen zwischen Termen dieses Vokabulars her.

„Nach DIN 1463 ist ein Thesaurus eine geordnete Zusammenstellung von Begriffen mit ihren Beziehungen.“ D.h., ein Thesaurus ist eine alphabetisch und systematisch geordnete Sammlung von Begriffen eines bestimmten Fachbereichs und dessen semantische Beziehungen.

---

[77] Vgl. Renz/Franke (2003), S. 5-6
[78] Vgl. Ackermann (2002), S. 35
[79] Vgl. Ferber (2003), S. 41
[80] Ferber (2003), S. 54

4.4.3.3 Stoppwortlisten

Stoppwörter sind häufig auftretende Terme, bei denen es sich in der Regel um Funktionswörter wie Artikel, Konjunktionen, Präpositionen und Hilfsverben handelt, aber auch Begriffe, deren Informationsgehalt nur gering ist, zählen dazu. Stoppwortlisten bestehen im Normalfall aus einigen hundert Begriffen.[81]

Die Grundlage der Entscheidung, welche Wörter als unwichtig und damit als Stoppwörter einzustufen sind, bildet die Statistik. Die zu stellende Frage lautet: Welche Wörter kommen auffallend häufig in nahezu allen Texten vor?

4.4.3.4 Merkmalsextraktion

Die einfachste Form der Merkmalsextraktion ist das Zerlegen eines Textes in Wörter, auch „Tokenisierung“ genannt. Nebenbei kann dieser Prozess Satzgrenzen erkennen und Normalisierungen durchführen wie z.B. die Vereinheitlichung von Schreibvarianten. Direkte Anwendungen der Merkmalsextraktion sind die Hervorhebung (z.B. Unterstreichung) wichtiger Ausdrücke eines Textes sowie die Extraktion von repräsentativen Wörtern (Schlüsselwortextraktion) und Sätzen (automatische Textzusammenfassung).[82]

4.4.3.5 Klassifikation

Wesentliches Ziel der Textklassifikation ist es, die inhaltliche Erschließung von großen Textmengen zu automatisieren. Hierbei wird zwischen lernenden und nichtlernenden Systemen unterschieden.

4.4.3.5.1 Lernende, überwachte Systeme

Lernende Systeme berechnen die Klassifikationsobjekte automatisch anhand von Trainingsbeispielen. Bei Trainingsbeispielen handelt es sich um eine möglichst repräsentative Auswahl bereits eindeutiger Texte. Durch den so genannten Klassifikator oder Kategorisierer werden die Objekte dabei vorgegebenen Kategorien zugeordnet. So nennt sich das Verfahren auch Kategorisierung.

Das Training besteht im Wesentlichen aus zwei Schritten: der Merkmalsauswahl und der Berechnung des Klassifikationsobjekts. Die Merkmalsauswahl dient dazu, die eindeutigen Merkmale der Trainingstexte auf die wichtigsten zu reduzieren, um die Schwierigkeiten und die Komplexität bei den Berechnungen der Objekte gering zu halten. So werden im Endeffekt für jede einzelne Klasse die dazu gehörigen Merkmale erkannt. Auf Grund dessen kann anschließend eine schnellere Zuordnung der Trainingsdokumente zu den ihnen entsprechenden Klassen erfolgen. Der Klassifikator hat somit gelernt.

---

[81] Vgl. Ackermann (2002), S. 26-27

[82] Vgl. Dörre et al. (2001), S. 435-436

Der Trainingsprozess selber wird mehrere Male wiederholt, und die dabei aufkommenden Fehler werden analysiert. Die Fehleranalyse führt dann zu Verbesserungsmaßnahmen des Systems. Eine derartige Verbesserung des Modells entspricht einem so genannten Lernschritt. Auf Grund dessen lernt das System nach und nach, die Trainingsdaten besser nachzubilden.

Nach Abschluss des Trainingsprozesses kann der Klassifikator dann auf die unbekannten, nichtklassifizierten Datensätze der Dokumentenkollektion angewendet werden, was letztendlich auch dem eigentlichen Zweck entspricht. Auf diese Weise werden die Strukturen der Trainingsdaten für neue, unbekannte Daten verallgemeinert.[83]

Abbildung 19 zeigt den Prozess dazu nochmals visuell auf.

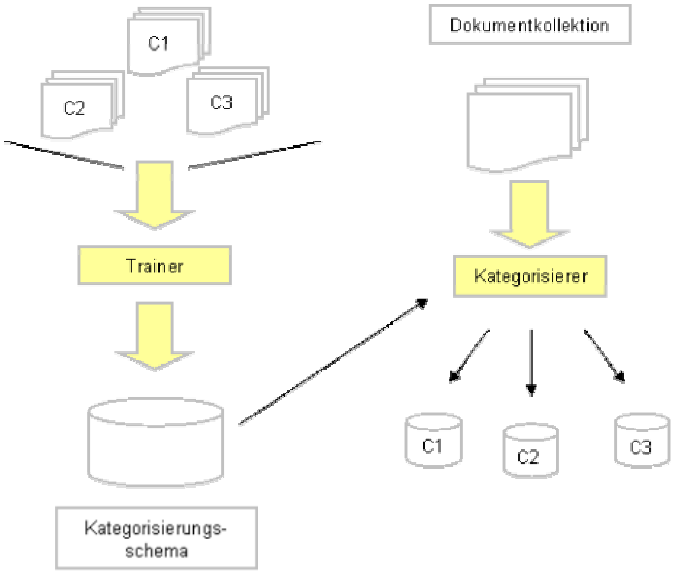


Abbildung 19: Vorgang der Kategorisierung[84]

Bei der Kategorisierung können zahlreiche Verfahren unterschieden werden. Zu den bekanntesten zählen der Rocchio-Algorithmus, die Support-Vector-Machine/SVM und die Entscheidungsbaumverfahren.

[83] Vgl. Krahl et al. (1998), S. 61-62
[84] Quelle: Dörre et al. (2001), S. 438

4.4.3.5.2 Nichtlernende, unüberwachte Systeme

Bei dieser zweiten Variante werden für die einzelnen Klassen üblicherweise durch menschliche Experten Regeln definiert. Man spricht hierbei auch von Clustering.

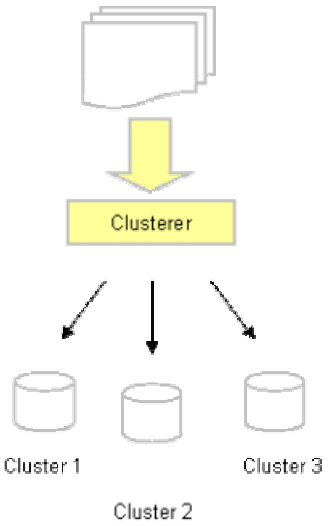


Abbildung 20: Vorgang des Clustering [85]

Clustering bedeutet, wie in Abbildung 20 dargestellt, dass eine Dokumentenkollektion in Gruppen oder Teilmengen (die so genannten Cluster) aufgeteilt wird. Ziel ist es, Gruppen zu bilden, deren Objekte thematisch möglichst ähnlich sind und sich möglichst stark von Objekten anderer Gruppen unterscheiden. In der Regel wird eine Ähnlichkeitsstruktur zwischen den Clustern ermittelt, die ein Distanzmaß zwischen den Clustern ausdrückt. Das dafür zuständige Vektorraummodell ist ein mathematisch einfaches und gut handhabbares Modell. Es geht von einem Vektorraum aus, der für jeden Term eine eigene Koordinatenachse besitzt. Jedes Dokument kann als Punkt in diesem n-dimensionalen Vektorraum angesehen werden und wird durch einen Vektor in dem Raum dargestellt. Die Terme, die in dem Dokument auftauchen, werden dabei gewichtet. Das kann zum einen dadurch geschehen, dass den Termen von vornherein unterschiedliche Gewichtungen gegeben werden, zum anderen können aus der Art des Auftretens in den Dokumenten unterschiedliche Gewichtungen abgeleitet werden.
Das Vektorraummodell ist besonders deshalb für viele Anwendungen attraktiv, weil Ähnlichkeiten zwischen Dokumenten oder zwischen Dokumenten und einer Anfrage sehr einfach als Abstand oder Winkel zwischen den Vektoren bestimmt werden können. Ein kleiner Abstand entspricht dann einer großen Ähnlichkeit und umgekehrt.

Analog dazu sollen in einem Information Retrieval System genau diejenigen Dokumente eine hohe Ähnlichkeit zum Anfragevektor erhalten, die auch für die Anfrage tatsächlich relevant sind. Neben den Klassifizierungsvorgängen, können die Cluster-Analysen auch Suchvorgänge und somit Recherchen mit Hilfe von Beispieldokumenten unterstützen. Das Ziel dieser Verfahren besteht darin, das Ergebnis zu optimieren, damit die

[85] Quelle: Dörre et al. (2001), S. 437

Ähnlichkeitsstruktur innerhalb der Cluster möglichst minimal und die Distanz zwischen verschiedenen Clustern möglichst maximal ist.[86]

4.4.4 Produktbeispiel „Readware IP-Server“

Im Rahmen einer möglichen Lösungsvariante für ein Text-Mining-Verfahren bestand für IPM/C bereits Interesse an einem Produkt der Firma Readware.

Das bekannte Produkt „Readware IP-Server“, unterstützt ebenfalls das Auffinden und Klassifizieren von Informationen in Intranet oder Internet. In weiten Anwendungsbereichen findet die Lösung ihren Einsatz und verfügt dabei über vielfältige Anpassungsmöglichkeiten an die unterschiedlichen Anforderungen der spezifischen Unternehmensbereiche. Möglichkeiten der Anpassung bestehen hinsichtlich der Plattformen, der Oberflächen oder des Speicherorts.

Mittels der automatischen Klassifikation können Dokumenteninhalte erschlossen und die Dokumente selber verschlagwortet werden. Über eine semantische Suche wird das Auffinden von Dokumenten ermöglicht. Das System ist dabei sprachunabhängig und besitzt ferner die Fähigkeit, mit fremdsprachigen Informationen umzugehen.

Da es sich jedoch um ein semantisches Verfahren handelt, bringt das weitere Anpassungen für die Patentabteilung mit sich. Unter anderem müssten die bestehenden Rechercheprofile in regelmäßigen Abständen aktualisiert werden - und das ist unter dem Aufwandsaspekt für IPM/C nicht tragbar.

Folglich fällt die Wahl auf die Realisierung einer statistischen Textanalyse. Recherchen ergeben, dass eine interne Forschungseinheit bereits eine Lösungsalternative entwickelt hat und auf dem Gebiet bewandert ist.

Letzten Endes entscheidet man sich zugunsten des internen Verfahrens „What’s Related“, das im Folgenden für die Anforderungserfüllung seitens IPM/C auf das Genaueste hin untersucht werden soll.

[86] Vgl. Dörre et al. (2001), S. 436-440

5 Das System „What's Related“

In diesem Kapitel wird das Text-Mining-Verfahren What's Related vorgestellt. Dabei werden sowohl die Arbeitsweise des Systems an sich als auch die Möglichkeiten, die durch What's Related geboten werden, näher beleuchtet.

5.1 Arbeitsweise

Das System What's Related ist ein modular aufgebautes Text-Mining-System, das seit den 90er Jahren am DaimlerChrysler-Forschungszentrum Ulm in der Abteilung Department of Data Mining Solutions, RIC/AM, Arbeitsgruppe Text-Mining, entwickelt und mittlerweile für verschiedene Anwendungen eingesetzt wird. Das System vereinigt dabei verschiedene Techniken und Methoden aus dem Information Retrieval, der statistischen Sprachverarbeitung und der Mustererkennung. Damit kann eine effiziente Repräsentation textueller Daten, eine Ermittlung von Ähnlichkeiten zwischen Dokumenten und eine lernende- oder nichtlernende Klassifikation erfolgen. Das System ist dabei sprachunabhängig.

What's Related hilft, die bestehende Dokumentenkollektion zu organisieren und zu strukturieren. Dabei werden vor allem zwei Ziele verfolgt: Zum einen soll die Dokumentenkollektion effektiv durchsucht werden, indem ausgehend von einem interessanten Dokument thematisch verwandte und somit ähnliche Dokumente gefunden werden. Und zum anderen soll die Wartung und Pflege der Dokumentensammlung unterstützt werden, da durch das System Vorschläge zur Strukturierung und Kategorisierung gemacht werden.

Das System kann in einen Teil zur Offline- und in einen Teil zur Online-Berechnung unterteilt werden. Offline bedeutet, dass die zugrunde liegenden Berechnungen auf einer Dokumentenkollektion nur einmal ausgeführt zu werden brauchen. Dieser Vorgang läuft auf einem separaten Rechner, dem Server, ab. Der Nutzer gibt eine Anfrage von seinem Arbeitsplatz aus in eine Recherchemaske ein, und dazu werden über das Intranet die Online-Berechnungen ausgeführt. Dieser Vorgang muss dabei für jede Anfrage neu ausgeführt werden.

5.1.1 Offline-Berechnung

Die Arbeitsschritte, die im Rahmen der Offline-Berechnung durchlaufen werden, sollen im folgenden Abschnitt der Reihenfolge nach näher betrachtet werden.

1. Zuerst erfolgt die Erfassung der Kollektion.
2. Als zweites erfolgt das Einlesen und Normieren der Texte. Zunächst entfernt das Verfahren alle Formatierungen und Satzzeichen aus den Dokumenten, damit eine lange Kette von Wörtern übrig bleibt. Diese Wörter sortiert das Programm dann nach der Häufigkeit ihres Auftretens im Text. Einige Wörter tauchen extrem

häufig, andere Wörter hingegen treten eher selten auf. Das Programm entfernt nun alle Wörter in den beiden Extrembereichen der Häufigkeitsverteilung; man spricht dabei von den Stoppwörtern.

Nachdem die Stoppwörter alle entfernt sind, wird auf Grundlage der verbleibenden Textteile die Zerlegung von Wortformen durchgeführt. Das Hauptziel ist die Gewinnung einer verringerten Anzahl von für die Klassifikation aussagekräftigen Schlagwörtern. Bei der Zerlegung wird überprüft, ob kürzere Wortformen in längeren Wortformen enthalten sind. Zutreffendenfalls wird eine längere Wortform in die darin enthaltene kürzere Wortform und in mindestens einen verbleibenden Wortteil zerlegt. Der Abgleich wird dann solange fortgeführt, bis keine weitere Zerlegung mehr möglich ist.

3. Erzeugen der Texel-Vektoren[87] und eines Texel-Lexikons[88]. Beide können mit verschiedenen Textgenerierungsmethoden erzeugt werden. Alternativ kann für die Generierung der Texel-Vektoren bereits ein vorhandenes Texel-Lexikon verwendet werden. Dann werden nur die Terme aus dem Dokument indiziert, die bereits im Lexikon enthalten sind.
4. Merkmalsreduktion durch Merkmalsselektion oder Stemming.
5. Berechnen der gewichteten Merkmalsvektoren.
6. Berechnung der Ähnlichkeitsmatrix und ggf. Kombination verschiedener Ähnlichkeitsmatrizen und Speicherung in der Ähnlichkeitsmatrix. In ihr kann für jeden aktuellen Text die Ähnlichkeit zu den anderen Texten abgelesen werden, auf Grund derer die thematisch verwandten Dokumente ausgegeben werden.
7. Optional: Training eines Klassifikators.
8. Optional: Clustering der Ähnlichkeitsmatrix.
9. Optional: Berechnung von Schlüsselwörtern und Textzusammenfassungen.

5.1.2 Online-Berechnung

Bei der Online-Berechnung hingegen erfolgt die Berechnung einer Anfrage oder eines Anfragedokuments zur inhaltsbasierten Suche. Dazu sind folgende Schritte notwendig:

1. Berechnung des Texel- und des gewichteten Merkmalsvektors für den Anfragetext auf Basis des vorhandenen Texel- und/oder Merkmalslexikons.
2. Berechnung der ähnlichsten Dokumente.
3. Generierung der Ergebnisliste.

---

[87] Ein Texel ist eine Zeichenkette, eine Ziffer oder ein Buchstabe und repräsentiert einen Indexterm gemeinsam mit einer eindeutigen Nummer sowie seiner Häufigkeit im Bestand.

[88] Ein Texel-Lexikon besteht demnach aus einzelnen Texel-Elementen.

4. Klassifikation neuer nichtklassifizierter Dokumente. Dazu werden die Merkmalsvektoren für die Dokumente berechnet und diese dann durch den trainierten Klassifikator zugeordnet.[89]

Die nachfolgende Abbildung soll die gerade beschriebene Arbeitsweise nochmals zusammenfassen:

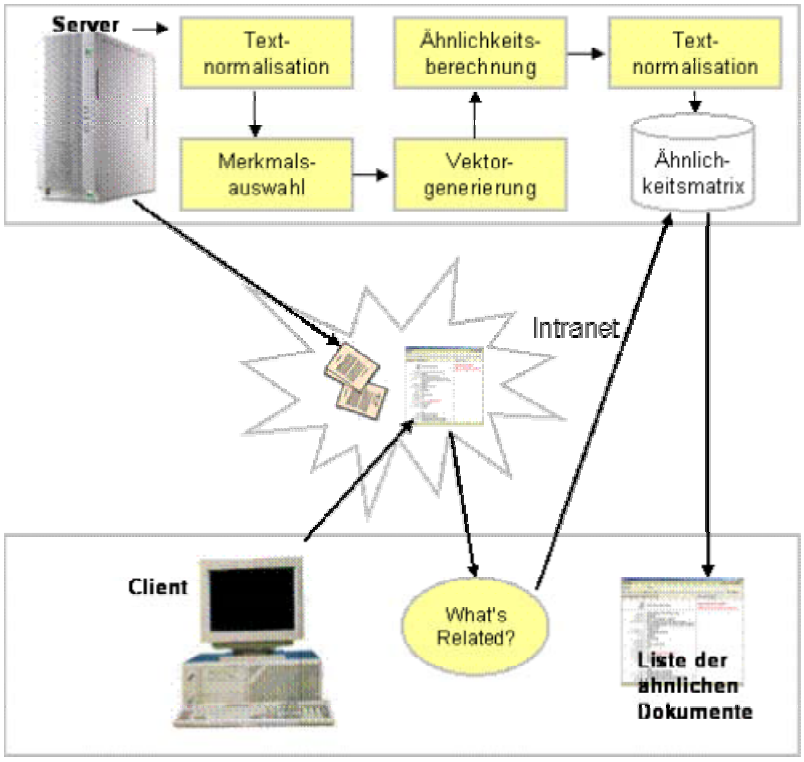


Abbildung 21: Offline-Berechnung und Online-Präsentation[90]

Für jedes bestehende Dokument werden schließlich nach Überprüfung der Ähnlichkeitsmatrix alle thematisch ähnlichen Dokumente angezeigt. Das WR-System arbeitet dabei wie eine Suchmaschine: Stichwörter müssen zur Suche eingegeben werden, und eine Liste mit signifikanten Texten wird ausgegeben. In dieser Liste kommen entweder Dokumente vor, in denen der gesuchte Begriff enthalten ist oder die dem gesuchten Begriff thematisch ähnlich sind. Die Ergebnisliste ist dabei nach der Relevanz, also der Gewichtung der gefundenen Suchergebnisse, geordnet. Als Grundlage liegt eine Datenbank dahinter, in der bei jeder Suchanfrage nach dem Suchbegriff des Nutzers gesucht wird. Die Datenbank enthält dabei die Erstveröffentlichungsschriften in Kurzauszügen, das heißt die Hauptansprüche bzw. Abstracts aus Deutschen Offenlegungsschriften, US-Patentschriften und Europäischen sowie Internationalen Anmeldungen.

[89] Vgl. Renz (1995), S. 2-3
[90] Quelle: Angelehnt an Bohnacker et al. (2002), S. 3

5.2 Arbeiten mit dem System

Die Web-Schnittstelle (im Folgenden nur noch Interface) der ersten Testversion für IPM/C steht auf einem Arbeitsplatzrechner der Ulmer Forschungsabteilung zur Verfügung. Sie ist provisorisch, reicht zum Testen der Funktionalitäten jedoch aus.

Das angebotene Interface besteht aus zwei Teilen: Im oberen Bereich wird die Suchanfrage eingegeben, im unteren Bereich werden die Suchergebnisse aufgelistet.

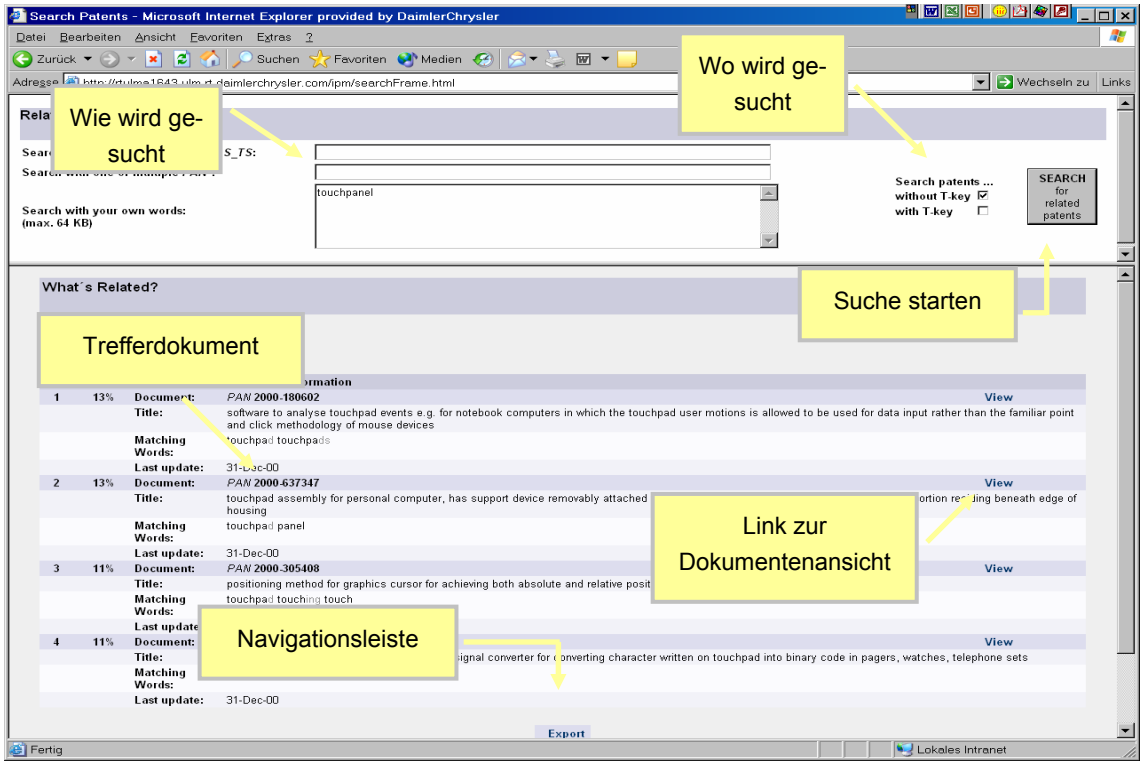


Abbildung 22: Webinterface des Systems What's Related

5.2.1 Wie kann gesucht werden?

Es gibt drei Arten von Suchanfragen, die beliebig miteinander kombiniert werden können.

a.) Search with one or multiple T-Keys S_TS

Hier können ein oder mehrere T-Schlüssel eingeben werden. Groß- und Kleinschreibung spielt dabei keine Rolle.

b.) Search with one or multiple PANs

Hier können ein oder mehrere Patentdokumente als Referenz für die Suche eingegeben werden. Als eindeutiger Schlüssel für Patente wird die PAN (Primary Accession Number) verwendet. Diese Nummer ist für die jeweilige Patentschrift eindeutig.

c.) Search with your own words

Hier kann freier Suchtext eingegeben werden. Das können entweder einzelne Wörter bis hin zum kompletten Freitext sein. Synonyme werden allerdings nicht beachtet. Die Suchargumente selber können nur komplett angegeben werden, eine Trunkierung ist nicht möglich. Des Weiteren müssen Suchargumente nicht durch Operatoren miteinander verglichen werden. Es reicht die Eingabe eines Trennzeichens in Form eines Kommas oder eines Leerzeichens.

5.2.2 Wo wird gesucht?

Die Dokumentenkollektion besteht aus zwei Teilen:

- Dokumente ohne T-Schlüssel, im Folgenden *nichtklassifizierte Dokumente* (→ Checkbox „Search patents without T-key“)
- Dokumente mit T-Schlüssel, im Folgenden bereits *klassifizierte Dokumente* (→ Checkbox „Search patents with T-key“)

Über die beiden Checkboxen „Search patents without T-key“ und „Search patents with T-key“ wird festgelegt, auf welchem Teil der Dokumentenkollektion gesucht wird.

5.2.3 Wie sehen die Ergebnisse aus?

Die Suche wird gestartet durch Anklicken der Schaltfläche „*Search for related patents*“ oder durch Drücken der Taste <ENTER> nach Eingabe einer Anfrage. Eine Anfrage kann immer wieder modifiziert und erneut ausgeführt werden.

Ergebnis der Suche ist eine Liste von Treffern, absteigend sortiert nach ihrer Ähnlichkeit zu der Anfrage. Dabei kann festgelegt werden, bis zu welchem Schwellwert die Patentdokumente in die Ergebnisliste aufgenommen werden. Präsentiert werden zehn Treffer auf einer Seite. Zu den weiteren Dokumenten kommt man über eine Navigationsleiste am Ende der Ergebnisliste. Sowohl die maximale Anzahl der Treffer als auch die Treffer pro Seite sind konfigurierbar. Über einen Export-Link besteht die Möglichkeit, alle Treffer auf einmal anzeigen zu lassen und die Daten durch Copy&Paste in einem Editor weiterzubearbeiten.

Ähnlichkeiten werden sowohl auf der Ebene von Wörtern als auch auf der Ebene von Teilwörtern der Länge vier, so genannte Quadgramme, berechnet. Fast alle Dokumente haben ein oder mehrere Quadgramme gemeinsam und damit untereinander eine Ähnlichkeit größer null.

Für einen Treffer werden folgende Daten angezeigt:

- Similarity

Dabei handelt es sich um ein Maß für die Ähnlichkeit von Anfrage und Trefferdokument im Bereich von 10 bis 100 Prozent. 100 Prozent bedeutet, dass Anfrage und Treffer Duplikate sind. Bei Anfrageart „Search with one or multiple PANs“ sollte das Anfragedokument selbst mit 100 Prozent gefunden werden.

- Rank

Das Ranking (oder die Rangfolge) der Trefferdokumente ergibt sich aus den Ähnlichkeitswerten. Je ähnlicher und relevanter ein Dokument zu einer Anfrage ist, desto höher steht es in der Ergebnisliste.

- Die Werte der Datenbankfelder PAN, S_TS und S_TEXT_D

Mit diesen Werten wird das Dokument identifiziert. PAN benennt das Trefferdokument mit seiner Primary Accession Number, S_TS zeigt den dazugehörigen Technologieschlüssel auf, und S_TEXT_D schreibt den T-Schlüssel als Text aus (z.B. T-Schlüssel Fußgängerschutz). Das Feld S_TS ist nur bei dem Teil der Dokumente vorhanden, die bereits klassifiziert sind. Darüber hinaus ist S_TEXT_D nur in der Test-Version vorhanden, in der realen Anwendung später nicht mehr.

- Title

Hier wird der Text des Datenbankfelds TI1 dargestellt und entspricht dem Titel des Dokuments.

- Matching Words

Hierbei handelt es sich um eine Liste der Wörter im Trefferdokument, die hauptsächlich für die Ähnlichkeit zur Anfrage verantwortlich sind. Die Liste ist absteigend sortiert nach dem Gewicht, das diese Wörter zur Gesamtähnlichkeit beitragen. Das Gewicht wird angezeigt, wenn man den Mauszeiger auf ein *matching word* setzt. Es ist möglich, dass ein Wort nicht als ganzes „match“, sondern nur dessen Wortteile, die Quadgramme, in die Gewichtung eingehen.

- Last update

Normalerweise wird hier das Datum der letzten Modifikation des Dokuments dargestellt. Da dieses Datum momentan nicht verfügbar ist, wird das Datum angezeigt, an dem das Dokument in das System eingetragen wurde.

- View

Damit wird ein Link bezeichnet, der eine Ansicht darstellt, innerhalb derer der Dokumenttext angezeigt wird. Dabei werden alle *matching words* bzw. Wortteile der *matching words* farbig markiert. Die folgende Abbildung zeigt die Darstellung der *matching words* auf.

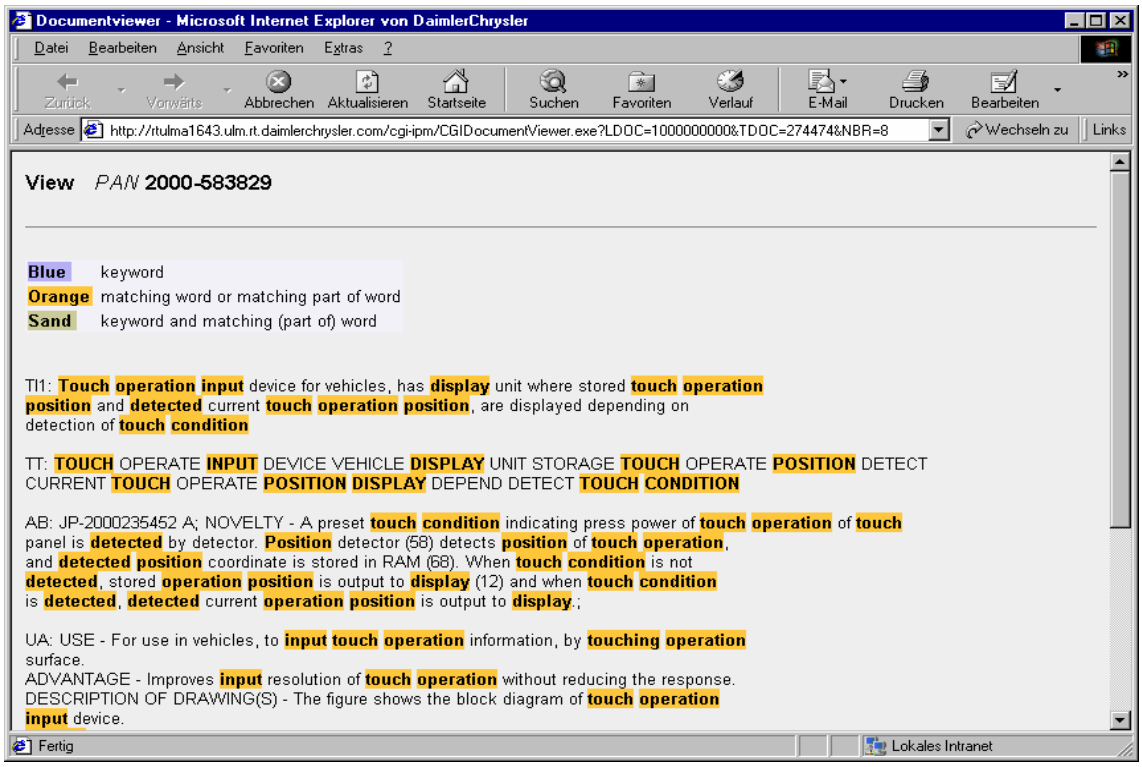


Abbildung 23: Darstellung der *matching words* in der View-Ansicht

5.2.4 Vorgehensweise beim Suchen

Es ist in der Regel nicht sinnvoll, die komplette Ergebnisliste durchzuschauen. Da davon ausgegangen wird, dass das Ranking der Dokumente sinnvoll ist, sollte der Anwender die Liste immer von oben nach unten durchgehen und dabei gedanklich seine eigenen Abbruchkriterien anwenden. Diese Abbruchkriterien können sein:

a.) Es werden genügend relevante Dokumente gefunden.

b.) Es besteht ein signifikanter Sprung im Similarity-Wert: Haben beispielsweise die ersten fünf Treffer Ähnlichkeitswerte zwischen 50 und 60 Prozent und der sechste Treffer hat nur noch einen Wert von 20 Prozent, dann kann davon ausgegangen werden, dass der sechste Treffer weniger relevant ist als die ersten fünf. Allerdings kann es auch sein, dass ein Treffer mit 20 Prozent an erster Stelle steht und genau das darstellt, wonach der Anwender sucht.

c.) Mehrere Dokumente nacheinander sind nicht relevant: In diesem besonderen Fall kann davon ausgegangen werden, dass auch die folgenden Dokumente nicht relevant sind. Die Relevanz eines Dokuments kann man meist am schnellsten an den *matching words* und dem Titel erkennen.

Werden nicht ausreichend Dokumente gefunden, sollte die Anfrage modifiziert werden.

5.2.5 Welche Kombinationen sind sinnvoll?

Das Hauptziel des WR-Systems ist es, in der Menge der nichtklassifizierten Dokumente (Dokumente ohne T-Schlüssel) diejenigen zu suchen, die am besten zu einem gegebenen T-Schlüssel passen. Dies ist die Kombination aus Anfrageart **(1)** „Search with one or multiple T-Keys S_TS“ und Dokumentenkollektion **(a)** „Search patents without T-key“, kurz **(1a).**

Aus der Systemarchitektur ergibt sich nun, dass die anderen Kombinationen ohne Zusatzaufwand ebenfalls möglich sind. Im Folgenden sind Bedeutung und Sinn der weiteren Kombinationen kurz dargestellt.

(1b) Eine Suche mit T-Schlüssel auf klassifizierten Dokumenten ist prinzipiell möglich. Sinnvoll ist eine solche Anfrage allenfalls als Hilfsmittel, um Hinweise auf falsch klassifizierte Dokumente zu erhalten.

(1ab) Eine Suche mit T-Schlüssel auf allen Dokumenten ist prinzipiell möglich, jedoch wenig sinnvoll.

(2a)(2b)(2ab) „Search with one or multiple PANs“ ist in jeder Kombination sinnvoll, da es hier nur darum geht, zu einem konkreten Dokument die Dokumente zu finden, die textuell ähnlich sind. Der T-Schlüssel spielt dabei keine Rolle.

(3a)(3b)(3ab) „Search with your own words“ ist in jeder Kombination sinnvoll. Es handelt sich dabei um eine Art „Suchmaschine“.

Es ist auch möglich, die verschiedenen Anfragearten miteinander zu kombinieren. Systemintern sind das dann getrennte Anfragen, deren Ergebnisse mit einem Mittelwert-Operator kombiniert werden.

Zur Verdeutlichung ein Beispiel:

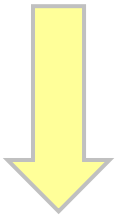
Die Abfrage mit dem T-Schlüssel ***„x“*** führt zu folgender Ergebnisliste:

Dokument	Ähnlichkeit
A	50%
B	50%
C	40%

Die Abfrage mit dem Suchwort ***„y“*** führt zu folgender Ergebnisliste:

Dokument	Ähnlichkeit
B	50%
D	30%
C	20%

Die Abfrage mit T-Schlüssel ***„x“*** und Suchwort ***„y“*** ergibt folgende Ergebnisliste:



Dokument	Ähnlichkeit
B	50%
C	30%
A	25%
D	15%

6 Evaluierung

Bevor die durchzuführenden Untersuchungen beschrieben werden, sollen konstruktive Methoden für eine Evaluierung dargelegt werden. Zu den Methoden zählen mögliche Werkzeuge, Richtlinien oder Standards.

6.1 Evaluationsmethodik

Die Methoden, die für die Evaluierung Anwendung finden sollen, werden nun dargestellt und deren Auswahl begründet.

6.1.1 Testmethoden

Gegenwärtig sind zwei Gruppen von Methoden bekannt: die Black-Box-Testmethode und die White-Box-Testmethode. Beim Black-Box-Test wird das Testobjekt als schwarzer Kasten angesehen. Das bedeutet, dass der Tester keine Informationen über die Struktur (z.B. Ablaufstruktur) des Testobjekts bei der Anwendung der Methode heranzieht. Lediglich die Leistungsbeschreibung des Testobjekts bildet die Basis zur Ableitung der Testvorgänge. Wichtig sind demnach nur die Eingaben und Ausgaben, was im System selber passiert, ist für die Evaluierung nicht weiter relevant.

Abbildung 24: Black-Box-Test[91]

Anhand konkreter Anwendungsfälle werden die Funktionen des Testobjekts identifiziert. Dazu wird für jede Funktion eine Anforderung erstellt. Anschließend werden die verschiedenen Testfälle durchgeführt, um zu zeigen, dass die Funktionen vorhanden und auch ausführbar sind. Die Testfälle sind dabei auf das Normalverhalten des Testobjekts ausgerichtet.

Der White-Box-Test, der für die Arbeit nicht weiter betrachtet wird, setzt voraus, dass die Struktur des Testobjekts bekannt ist und auch untersucht wird.[92]

[91] Quelle: Eigene Darstellung
[92] Vgl. Wallmüller (2001), S. 228-230

Des Weiteren wird die Testmethode des Cognitive Walkthrough ausgewählt, da sie aus der Fülle der verwendeten Methoden am ansprechendsten ist. Walkthrough ist eine aufgabenorientierte Inspektionsmethode und findet ohne Testpersonen statt. Lediglich ein Experte geht mit dem zu testenden System die einzelnen Schritte der Aufgabenerledigung durch und beurteilt, inwieweit das Produkt die Erledigung der Aufgabe unterstützt. Ein Cognitive Walkthrough bedeutet deshalb, dass „in den Schuhen des Benutzers durch die Software gelaufen wird."[93]

6.1.2 Effektivität

Effektivität ist die Fähigkeit eines Systems, relevante Dokumente aufzufinden und zu präsentieren. Nach Jiri Panyr[94] gilt sie als das Maß der Fähigkeit eines Systems, die Funktionen auszuführen, für die es vorgesehen ist. Folglich wird damit die Fähigkeit bezeichnet, dem Nutzer die Informationen nachzuweisen, die er sucht und die ihn bei seiner Aufgabenbewältigung unterstützen. Das System sollte demnach in der Lage sein, wiedergefundene Treffer nach der Wahrscheinlichkeit des positiven Relevanzurteils seitens des Benutzers zu ordnen, um eine optimale Effektivität zu gewährleisten. Damit wird die Relevanz als eine Relation zwischen einem Dokument und einem Benutzer in Bezug auf sein Informationsbedürfnis verstanden. Nach dieser Definition wird dann die Menge der relevanten Dokumente einfach der Menge derjenigen Dokumente gleichgesetzt, die das Informationsbedürfnis des Anwenders zufrieden stellt.

Das zentrale Problem bei der Beurteilung von Rechercheergebnissen besteht allerdings darin, dass die richtige Antwort bekannt sein muss, um die Antwort des Systems zu bewerten. Das heißt, es muss zum einen bekannt sein, welche Dokumente in der Datenbank stehen, und zum anderen, ob sie zu der Anfrage gehören, um schließlich bestimmen zu können, ob für eine Anfrage die richtigen Dokumente getroffen werden. Auch zu bestimmen, wann ein Dokument wirklich relevant ist, erweist sich als schwierig und ist ein zentraler Schwachpunkt, da immer nur pro Benutzer entschieden werden kann, wie relevant der Treffer wirklich ist. Das liegt daran, dass jeder Nutzer unterschiedliche Interessen und auch ein unterschiedliches Vorwissen hat.[95]

---

[93] Vgl. Erdmann (2002), S. 6

[94] Panyr (1986), S. 24-26

[95] Vgl. Haag (2002), S. 33-35

	Für eine Suchanfrage relevante Dokumente	Für eine Suchanfrage irrelevante Dokumente	Zeilensumme
Bei der Recherche gefundene Dokumente	a	b	a + b
Bei der Recherche nicht gefundene Dokumente	c	d	c + d
Spaltensumme	a + c	b + d	a + b + c + d

Tabelle 5: Rechercheergebnis als Häufigkeitstabelle oder Kontingenztafel[96]

Das Ergebnis einer Recherche kann wie in Tabelle 5 aufgegliedert werden. Dabei werden für die Felder folgende Bezeichnungen verwendet:

a: Zahl der Dokumente, die das System korrekt als relevant eingestuft hat

b: Zahl der Dokumente, die das System als relevant eingestuft hat, die aber tatsächlich nicht relevant sind

c: Zahl der Dokumente, die das System als nichtrelevant eingestuft hat, die aber tatsächlich relevant sind

d: Zahl der Dokumente, die das System korrekt als nichtrelevant eingestuft hat

6.1.2.1 Precision und Recall

Die Evaluierungsergebnisse von Textklassifikationssystemen werden meist in Recall (Vollständigkeit; eigentlich Abruf, deshalb wird im Folgenden auch von „der Recall“ gesprochen) und Precision (Genauigkeit) angegeben, um die Qualität hinsichtlich einer gestellten Anfrage zu messen.[97]

6.1.2.1.1 Precision

Bezogen auf die Informationsextraktion bezeichnet die Precision P den Anteil an korrekt gewonnenen Wissensobjekten (a) im Vergleich zu den insgesamt gefundenen Wissensobjekten ($a+b$). Eine hohe Precision bedeutet daher, dass fast alle gefundenen Wissensobjekte relevant sind.

Formel: $P = \frac{a}{a+b} * 100\%$

Der Wertebereich der Precision geht von null bis eins, bzw. von null bis hundert Prozent. Dabei wird versucht, den Wert zu maximieren.[98]

---

[96] Quelle: Gaus (2003), S. 216
[97] Vgl. Ackermann (2002), S. 97
[98] Vgl. Rolker (2002), S. 9

6.1.2.1.2 Recall

Der Recall R bezeichnet den Anteil der korrekt gewonnenen Wissensobjekte (a) im Vergleich zu den insgesamt gewinnbaren Wissensobjekten ($a+c$). Eine hohe Vollständigkeit bedeutet daher, dass fast alle relevanten Wissensobjekte gefunden werden.

Formel: $R = \frac{a}{a+c} * 100\%$

Der Wertebereich geht von null bis eins, bzw. von null bis hundert Prozent. Ein Recall von null wird für das schlechteste Ergebnis, eins für das bestmögliche vergeben. Bei der Patentdokumentation ist vor allem ein hoher Recall-Wert wichtig.[99]

6.1.2.1.3 Fazit

Die Idealwerte für Precision und Recall sind jeweils eins. Für die Precision ergibt sich dieser Idealwert, wenn die Antwortmenge z.B. nur ein einziges und relevantes Dokument enthält. Der Idealwert für den Recall ergibt sich, wenn die Antwortmenge z.B. gleich der Dokumentmenge ist. In der Regel wird die Antwortmenge zwischen diesen beiden Extremen liegen. Sinnvoll ist jedoch nur die Betrachtung beider Maße, da der Recall die Zahl der irrelevanten ausgegebenen Dokumente unberücksichtigt lässt und leicht auf das Maximum von eins gesetzt werden kann, indem alle im Dokumentenbestand vorhandenen Dokumente ausgegeben werden. In diesem Fall wäre der Precision-Wert allerdings sehr niedrig. Die alleinige Betrachtung der Precision wiederum würde nichts über die Vollständigkeit des Retrieval-Ergebnisses aussagen. Die Precision alleine könnte dadurch maximiert werden, dass nur sehr wenige Dokumente ausgegeben werden. Bei Suchanfragen mit einem Anspruch auf Vollständigkeit des Suchergebnisses wird ein hoher Recall angestrebt, so dass auch innerhalb dieser Arbeit ein größeres Augenmerk auf dieses Maß gerichtet wird.[100] Relevanz- und Vollzähligkeitsrate sind also voneinander abhängig und ergebenen erst zusammen einen Aussagewert.

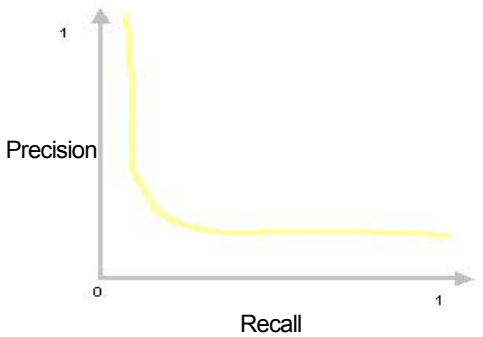


Abbildung 25: Precision und Recall[101]

[99] Vgl. Rolker (2002), S. 10
[100] Vgl. Rolker (2002), S. 11
[101] Quelle: Angelehnt an Ferber (2003)

6.1.2.2 Availability

Mit Hilfe einer so genannten Known-Item-Analyse wird nun versucht, das beschriebene Problem der Relevanz zu umgehen.

Bei dieser Analyse wird das zu untersuchende System mit unterschiedlichen Suchanfragen konfrontiert. Die Treffer, die herauskommen sollen, sind jedoch bekannt. Das Ziel einer derartigen Untersuchung besteht somit darin, möglichst alle der gesuchten Wissensobjekte bzw. davon so viele wie möglich zu treffen. Auf diese Weise lässt sich angeben, wie gut die Suchfunktionalität ist.

Abbildung 26: Availability einer Suchanfrage[102]

Nach Abbildung 26 bezeichnet die Availability A als Ergebnis der Known-Item-Analyse den Anteil der korrekt gewonnenen bekannten Wissensobjekte (Relevante Dokumente bei den tatsächlich erhaltenen Dokumenten; D_{gef}) im Vergleich zu den insgesamt gewinnbaren bekannten Wissenseinheiten (Relevante Dokumente; D). Eine hohe Availability bedeutet daher, dass fast alle bekannten und gesuchten Wissensobjekte extrahiert werden.

Gemessen wird damit die Verfügbarkeit oder Availability der Wissensobjekte. Der Messwert hat dabei zwei Ausprägungen; das so genannte Known-Item (das bekannte Dokument) wird gefunden oder nicht.[103]

Formel: $A = \frac{D_{gef}}{D} * 100\%$

Der Wertebereich geht von null bis eins, bzw. von null bis hundert Prozent. Eine Availability von null wird für das schlechteste Ergebnis, eins für das bestmögliche vergeben.

[102] Quelle: Eigene Darstellung
[103] Vgl. Dresel et al. (2001), S. 387

6.2 Durchführung der Untersuchung

Um festzustellen, wie effizient die in Kapitel 6.1 beschriebenen Methoden in der Praxis sind, sollen die Maßnahmen umgesetzt und angewandt werden. Die Ergebnisse dazu werden anhand von Beispielen veranschaulicht und interpretiert. Die Untersuchungsgegenstände werden dabei in folgender Reihenfolge erläutert:

1. Known-Item-Analyse
2. Schwellwertbestimmung
3. Einfluss der Internationalen Patentklassifikation

Vorab soll jedoch die Dokumentenkollektion, die als Grundlage für die durchgeführten Untersuchungen dient, vorgestellt werden.

6.2.1 Die Dokumentenkollektion

Die Dokumentenkollektion wird innerhalb von IPM/C thematisch ausgesucht, zusammengetragen und aufbereitet. Auf Grund der Tatsache, dass die Derwent-Daten in englischer Sprache verfasst und somit homogene Datensätze sind, werden für den ersten Testlauf des WR-Systems nur Derwent-Daten verwendet.

Abgelegt werden die Daten in einer Microsoft (im Folgenden nur noch MS) Access Datenbank. Anschließend wird die Kollektion an die Ulmer Forschungsabteilung zur Übertragung in das System übergeben.

Nach der Übertragung werden die Dokumente der Lernkollektion dazu verwendet, Referenzvektoren für die einzelnen T-Schlüssel zu berechnen. Dazu werden sie in verschiedene Kategorien eingeteilt, und der Kategorisierer erlernt anhand dieser Zuordnungen die Klassifikationsregeln. In der darauf folgenden Testphase werden die erlernten Regeln schließlich angewandt, um neue Texte in die bestehenden Kategorien einzuteilen.

Für den späteren Prozessablauf soll eine halbautomatische Zuordnung der Patentschutzrechte erfolgen. Dazu wird vom System überprüft, welche Klassifikationen bereits bestehen und wie die Dokumente diesen zugeordnet werden können. Halbautomatisch bedeutet dabei, dass die Administratoren die Vorschläge des Text-Mining-Verfahrens überprüfen und annehmen oder im Falle eines falschen Vorschlags die Dokumente richtig einordnen. Somit geht es im Endeffekt nur noch um die Wartung der Lernmenge.

Der derzeitige Datenbestand des WR-Systems ist dabei in eine Lern- und Testkollektion aufgeteilt.

Lernkollektion: Jeder Datensatz besitzt eine T-Schlüssel-Zuordnung	
DaimlerChrysler-Patente	2.055 Patente mit ca. 200 T-Schlüsseln aus verschiedenen Jahren
DaimlerChrysler- & Wettbewerber-Patente	4.998 Patente zum T-Schlüssel Laserschweißen ohne das Jahr 2000
	309 Patente zum T-Schlüssel Fußgängerschutz ohne das Jahr 2000
	45 Patente zum T-Schlüssel Touchpad ohne das Jahr 2000
Testkollektion: Kein Datensatz besitzt eine T-Schlüssel-Zuordnung	
DaimlerChrysler- & Wettbewerber-Patente	312.215 unbekannte Patente aus dem Jahr 2000
	894 bekannte Patente zum T-Schlüssel Laserschweißen aus dem Jahr 2000
	53 bekannte Patente zum T-Schlüssel Fußgängerschutz aus dem Jahr 2000
	12 bekannte Patente zum T-Schlüssel Touchpad aus dem Jahr 2000

Tabelle 6: Aufteilung der Lern- und Testkollektion[104]

Innerhalb der Lernkollektion besitzt jeder Datensatz einen Technologieschlüssel. Die Daten sind alle bekannt und kommen aus verschiedenen Jahrgängen. Bei den drei T-Schlüsseln Laserschweißen, Fußgängerschutz und Touchpad sind die bereits bekannten Zuordnungen des Jahres 2000 für spätere Testzwecke entnommen und nicht in das Lernsystem eingespeist worden. Sie werden aber als Teil der Testkollektion in WR integriert. Somit machen sie die Dokumentenmenge der Known-Items aus.

[104] Quelle: Eigene Darstellung
Im Verlauf der Arbeit stellt sich heraus, dass die bekannten Patentdokumente der Testkollektion für das jeweilige T-Schlüsselprofil nicht vollständig sind. Die Datenmenge wird deshalb im Weiteren bei relevanten Treffern um die Patentdokumente erweitert, die zusätzlich aufgefunden werden. Folglich ist diese Datenmenge später nicht mehr so klein wie zu Beginn.

6.2.2 Known-Item-Analyse

Im Rahmen der vorliegenden Arbeit wird eine Known-Item-Analyse durchgeführt und somit das Problem der Relevanzbeurteilung wirksam umgangen. Das Ziel der Untersuchung besteht darin, in mehreren Recherchen möglichst alle gesuchten Patentdokumente bzw. davon so viele wie möglich zu treffen. Das Besondere liegt darin, dass sämtliche Patentschutzrechte bereits bekannt sind. Auf diese Weise lässt sich angeben, wie gut die Suchfunktionalität ist, wie gut das WR-System aus seiner Referenzmenge gelernt hat und wie gut die Klassifikation erfolgt. Die Trefferquote zeigt damit die Verfügbarkeit des Datenbankbestands auf und legt im Endeffekt die Trefferqualität fest (siehe dazu Kapitel 6.1.2.2).

Vor der eigentlichen Durchführung werden noch einige Vorbereitungen für eine vereinfachte Bearbeitung der Trefferergebnisse und für eine bessere Vergleichbarkeit des Text-Mining-Verfahrens getroffen.

6.2.2.1 Vorbereitungen

Da die 2000er Dokumente der drei Technologieschlüssel Laserschweißen, Fußgängerschutz und Touchpad bereits lokal in einer MS Access Datenbank abgelegt sind und auch die Datenlieferung an die Data-Mining-Forschungsabteilung in diesem Format erfolgte, sollen die Auswertungen des Text-Mining-Verfahrens ebenfalls mit Hilfe einer dazu erstellten MS Access Datenbank durchgeführt werden. Dieser Vorgang wird im Folgenden erläutert.

Wenn eine Suchanfrage an das WR-System gestellt wird, werden alle Treffer in einer Ergebnisliste aufgeführt. Mittels eines Export-Links kann diese Treffermenge anschließend wie in Abbildung 27 dargestellt aufgelistet werden. Aufgeführt ist ein einziger String, der die Treffer-ID, die PAN des Patentdokuments, einen eventuell vorhandenen T-Schlüssel mit dessen Bezeichnung und den Ähnlichkeitswert aufzeigt.

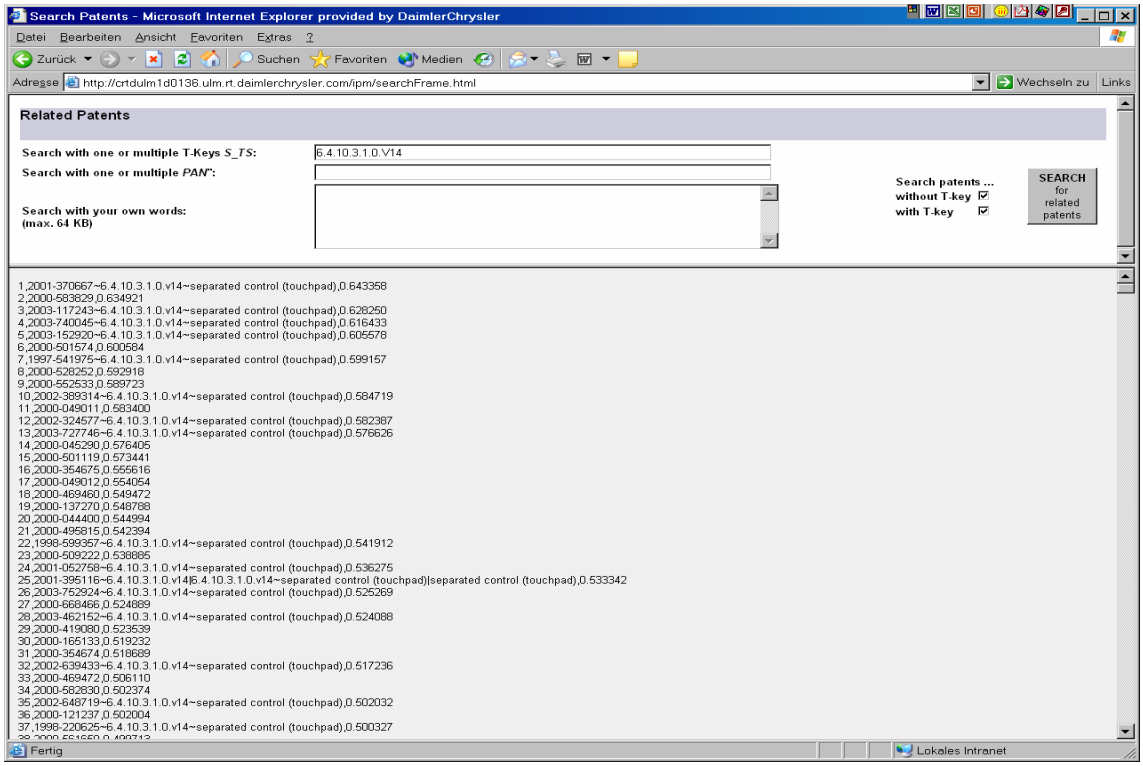
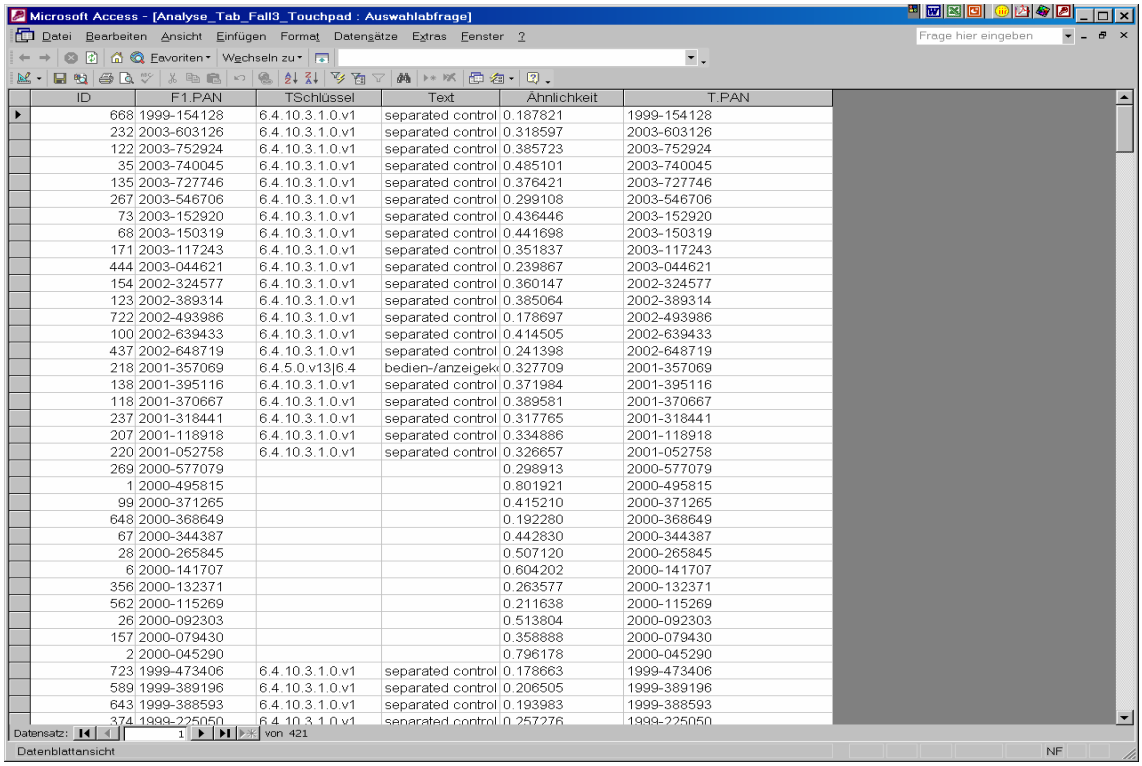


Abbildung 27: Trefferergebnis einer Suchanfrage

Die Daten werden anschließend mittels Copy&Paste in einen Texteditor und von dort aus durch ein MS Access Makro in die neu erstellte Datenbank importiert, wo sie schließlich in einer Tabelle gespeichert werden. Als nächster Schritt wird diese Tabelle mit der jeweiligen Originaltabelle der drei Technologieschlüssel Laserschweißen, Fußgängerschutz oder Touchpad verglichen. In den Originaltabellen befinden sich dabei sämtliche Informationen über das jeweilige Profil. Das sind unter anderem der Titel, die PAN oder der Anmelder. Durch diese Abfrage werden die übereinstimmenden Datensätze des Trefferergebnisses aus Abbildung 27 und der Originaltabelle ausgegeben.



ID	F1.PAN	TSchlüssel	Text	Ähnlichkeit	T.PAN
668	1999-154128	6.4.10.3.1.0.v1	separated control	0.187821	1999-154128
232	2003-603126	6.4.10.3.1.0.v1	separated control	0.318597	2003-603126
122	2003-752924	6.4.10.3.1.0.v1	separated control	0.385723	2003-752924
35	2003-740045	6.4.10.3.1.0.v1	separated control	0.485101	2003-740045
135	2003-727746	6.4.10.3.1.0.v1	separated control	0.376421	2003-727746
267	2003-546706	6.4.10.3.1.0.v1	separated control	0.299108	2003-546706
73	2003-152920	6.4.10.3.1.0.v1	separated control	0.436446	2003-152920
68	2003-150319	6.4.10.3.1.0.v1	separated control	0.441698	2003-150319
171	2003-117243	6.4.10.3.1.0.v1	separated control	0.351837	2003-117243
444	2003-044621	6.4.10.3.1.0.v1	separated control	0.239867	2003-044621
154	2002-324577	6.4.10.3.1.0.v1	separated control	0.360147	2002-324577
123	2002-389314	6.4.10.3.1.0.v1	separated control	0.385064	2002-389314
722	2002-493986	6.4.10.3.1.0.v1	separated control	0.178697	2002-493986
100	2002-639433	6.4.10.3.1.0.v1	separated control	0.414505	2002-639433
437	2002-648719	6.4.10.3.1.0.v1	separated control	0.241398	2002-648719
218	2001-357069	6.4.5.0.v13|6.4	bedien-/anzeigekı	0.327709	2001-357069
138	2001-395116	6.4.10.3.1.0.v1	separated control	0.371984	2001-395116
118	2001-370667	6.4.10.3.1.0.v1	separated control	0.389581	2001-370667
237	2001-318441	6.4.10.3.1.0.v1	separated control	0.317765	2001-318441
207	2001-118918	6.4.10.3.1.0.v1	separated control	0.334886	2001-118918
220	2001-052758	6.4.10.3.1.0.v1	separated control	0.326657	2001-052758
269	2000-577079			0.298913	2000-577079
1	2000-495815			0.801921	2000-495815
99	2000-371265			0.415210	2000-371265
648	2000-368649			0.192280	2000-368649
67	2000-344387			0.442830	2000-344387
28	2000-265845			0.507120	2000-265845
6	2000-141707			0.604202	2000-141707
356	2000-132371			0.263577	2000-132371
562	2000-115269			0.211638	2000-115269
26	2000-092303			0.513804	2000-092303
157	2000-079430			0.358888	2000-079430
2	2000-045290			0.796178	2000-045290
723	1999-473406	6.4.10.3.1.0.v1	separated control	0.178663	1999-473406
589	1999-389196	6.4.10.3.1.0.v1	separated control	0.206505	1999-389196
643	1999-388593	6.4.10.3.1.0.v1	separated control	0.193983	1999-388593
374	1999-225050	6.4.10.3.1.0.v1	separated control	0.257276	1999-225050

Abbildung 28: Übereinstimmende Datensätze der Treffer und der Originaltabelle

In Abbildung 28 ist die Ausgabe der Abfrage dargestellt. Die ersten fünf Felder „ID“, „F1.PAN“, „TSchlüssel“, „Text“ und „Ähnlichkeit“ kommen aus der Trefferergebnismenge, und das letzte Feld „T.PAN“ kommt aus der Originaltabelle des jeweiligen T-Schlüssels. Die Datensätze werden dabei über die beiden PAN-Felder verglichen und bei einer Übereinstimmung in der Ergebnismenge ausgegeben. Deutlich erkennbar, dass bei einigen Treffern bereits T-Schüssel zugeordnet sind, bei einigen jedoch nicht.

Die Aufgabe der Datenbank besteht letztendlich darin, die gewonnene Treffermenge aus der Suchanfrage vergleichbar zu machen.

6.2.2.2 Vorgehen bei der Known-Item-Analyse

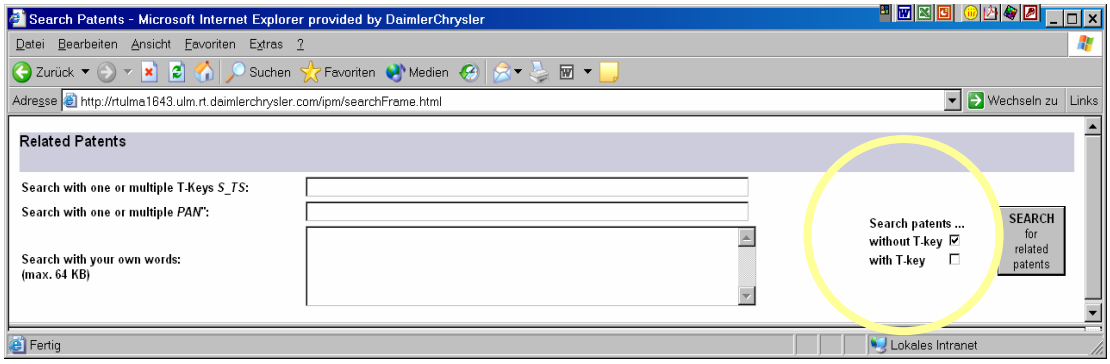
Zu Beginn werden die in Frage kommenden Suchanfragen für die Known-Item-Analyse in eine „Einfache Suche“ und „Kombinierte Suche“ untergliedert. Im Fall der einfachen Suche werden ausschließlich einzelne Suchbegriffe eingegeben, im Fall der kombinierten Suche werden kombinierte Möglichkeiten durchgespielt. Insgesamt erfolgen drei Recherchedurchgänge - entsprechend den drei T-Schlüsselprofilen Touchpad, Fußgängerschutz und Laserschweißen. Die Suchanfragen sind vom Prinzip her gleich, werden aber den Profilen angepasst.

Auf Grund der großen Zahl an Ergebnistabellen werden die Suchmöglichkeiten und Ergebnisse im Folgenden lediglich am Beispiel des Touchpad-Profils aufgezeigt.

6.2.2.2.1 Recherche nach Patentdokumenten ohne T-Schlüssel

Zieltreffer aus der Testkollektion

894 Patente TS Laserschweißen, **53** Patente TS Fußgängerschutz und **12** Patente TS Touchpad



Bei dieser Variante wird innerhalb der Testkollektion gesucht. Enthalten sind darin nur Datensätze des Derwent-Jahrgangs 2000. T-Schlüssel werden hierbei nicht berücksichtigt und tauchen in der Treffermenge nicht auf. Das Ziel dieser Recherchevariante besteht darin, von den entnommenen, lokal abgelegten und bereits bekannten Patentdokumenten der drei T-Schlüssel Laserschweißen, Fußgängerschutz und Touchpad so viele Patente wie möglich zu treffen. Als Treffer wird gezählt, wenn das bekannte Patent gefunden wird.

❖ Einfache Suche

1. Die erste Eingabemöglichkeit erfolgt mit dem T-Schlüssel 6.4.10.3.1.0.V14.

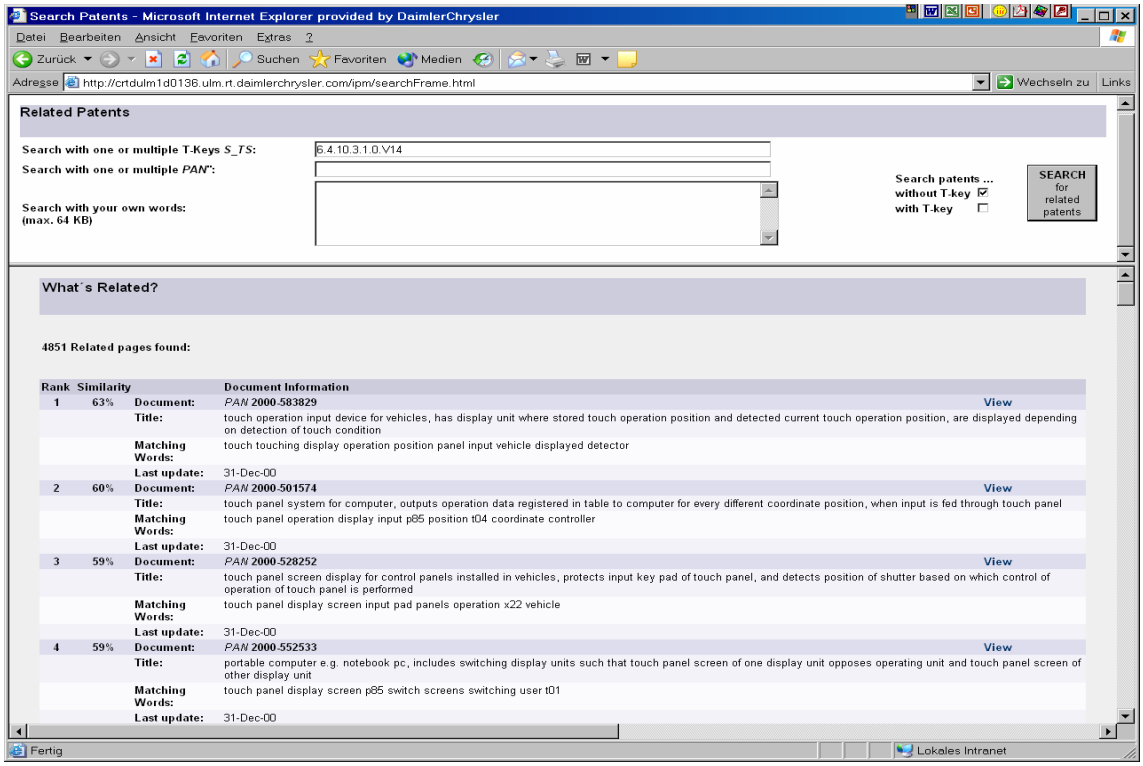


Abbildung 29: Suchanfrage bei Eingabe des Technologieschlüssels

2. Die zweite Eingabe erfolgt mit der Primary Accession Number (PAN) der Patentschutzrechte. Davon werden nacheinander fünf Stück ausgewählt und eingegeben. Die Beispieldokumente sind dem Touchpad-Profil eindeutig zuordnungsfähig und werden aus der Originaltabelle entnommen.

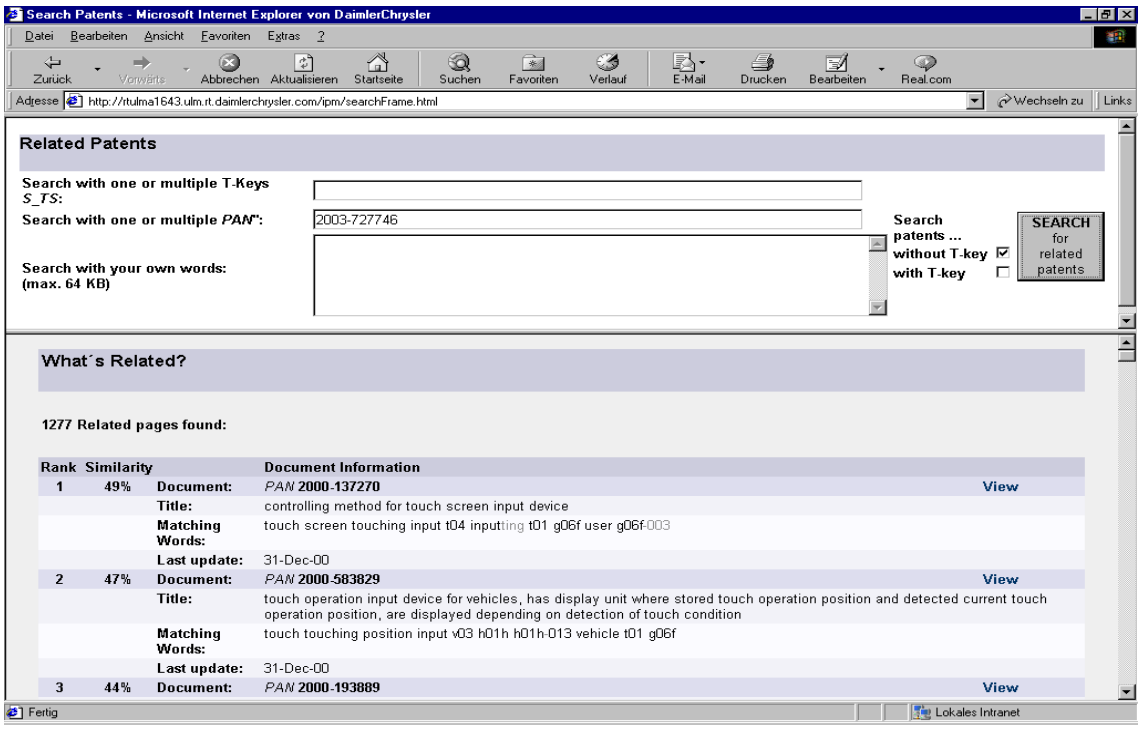


Abbildung 30: Suchanfrage bei Eingabe eines Beispieldokuments

3. Bei der letzten Variante werden freie Begriffe eingegeben, die dem Profil ebenfalls eindeutig zugeordnet werden können, wie z.B. „touch pad“ oder „touch“.

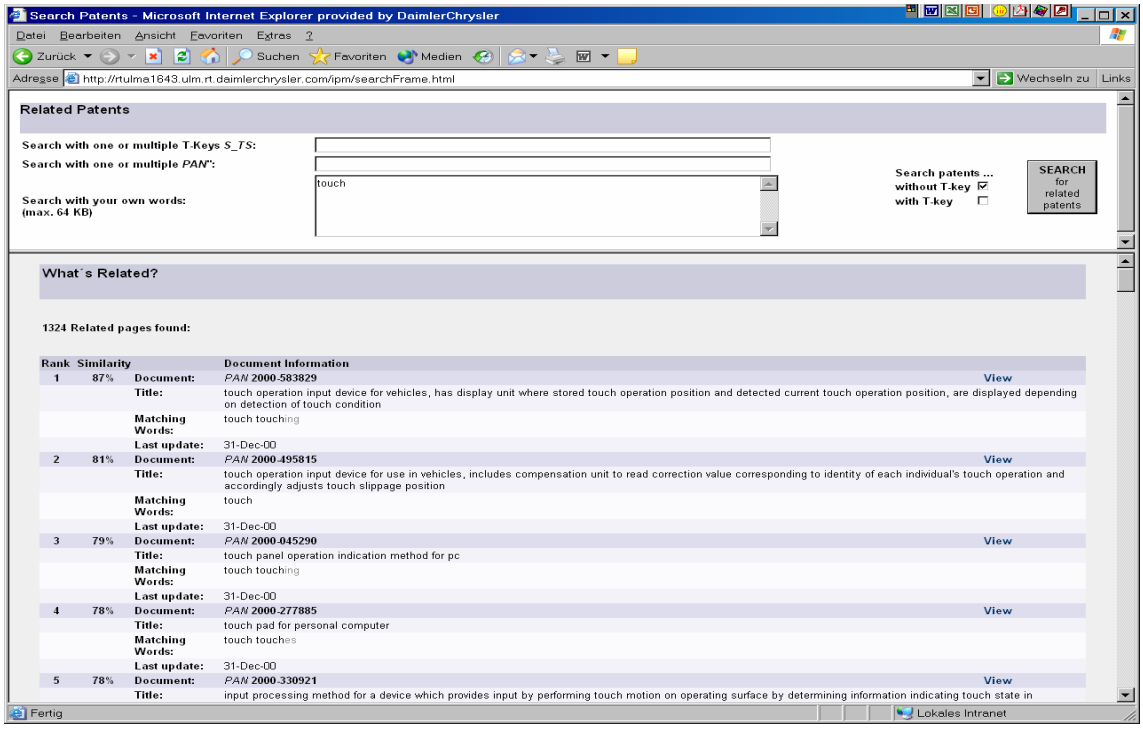


Abbildung 31: Suchanfrage bei Eingabe eines Schlagwortes

❖ Kombinierte Suche

1. Fall: Bei dieser Möglichkeit werden dieselben Begriffe der einfachen Suche übernommen und untereinander kombiniert.

2. Fall: Eine weitere Idee besteht darin, einzelne getroffene Patente der entnommenen 2000er Dokumente in die Suchanfrage aufzunehmen.

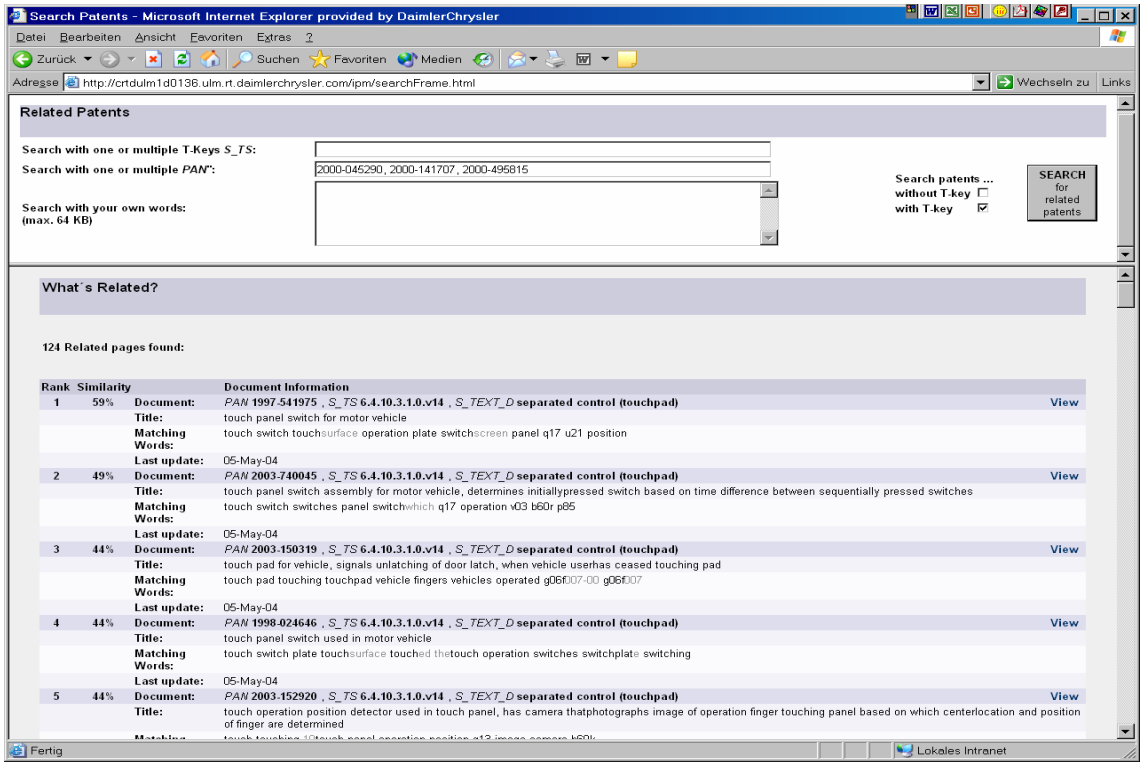
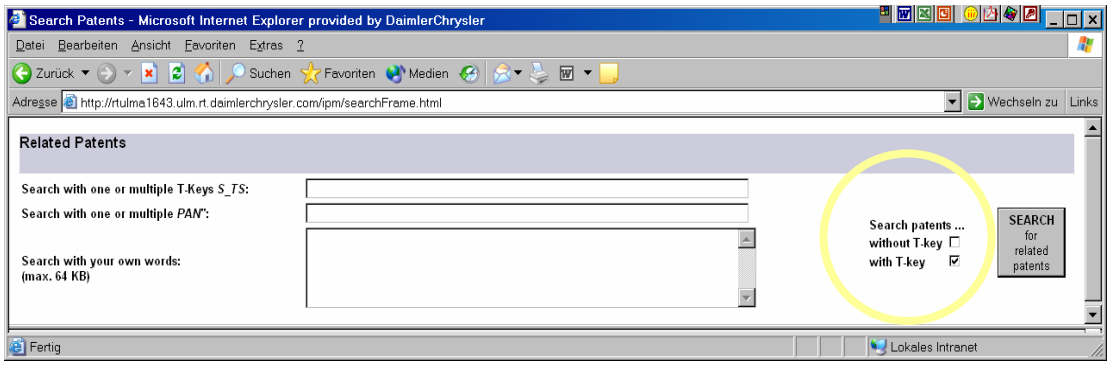


Abbildung 32: Suchanfrage bei Eingabe getroffener relevanter Beispieldokumente

6.2.2.2.2 Recherche nach Patentdokumenten mit T-Schlüssel

Zieltreffer aus der Lernkollektion

4.998 Patente TS Laserschweißen, **309** Patente TS Fußgängerschutz und **45** Patente TS Touchpad

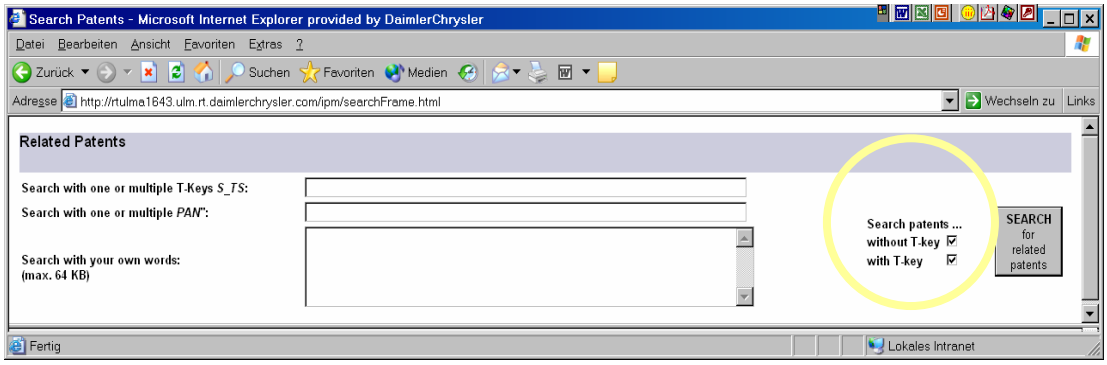


Innerhalb dieser Möglichkeit wird die gesamte Lernkollektion untersucht. Die Patentdokumente dieser Menge haben bereits einen T-Schlüssel zugeordnet. Aus diesem Grund sind auch keine Patente aus dem Jahr 2000 enthalten, sondern nur bereits

klassifizierte Dokumente. Das Ziel besteht darin, zu überprüfen wie gut die Klassifizierung funktioniert hat und wie viele Patentdokumente letztendlich aus der Lernmenge gefunden werden. Verfolgt wird dabei dasselbe Vorgehen wie in Kapitel 6.2.2.2.1.

6.2.2.2.3 Recherche nach Patentdokumenten mit oder ohne T-Schlüssel

Zieltreffer aus der Test- und Lernkollektion
5.892 Patente TS Laserschweißen, **362** Patente TS Fußgängerschutz und **57** Patente TS Touchpad



Die dritte Möglichkeit schließlich umfasst die gesamte Datenkollektion, das heißt, darin ist die gesamte Test- und Lernmenge enthalten. Das Ziel ist zu überprüfen, wie viele Patente insgesamt getroffen werden, egal ob in der Lernkollektion enthalten oder ob bereits neu zugeordnet. Auch hier erfolgt die Suche mit denselben Begriffen wie aus Kapitel 6.2.2.2.1.

6.2.2.3 Ergebnisse

Es werden allgemeingültige Ergebnisse beobachtet, die bei allen drei T-Schlüsselprofilen auftauchen.

Für die erste Eingabemöglichkeit mittels des Technologieschlüssels ist nur eine Eingabeart zulässig: Der T-Schlüssel selber. Eine Ebene höher oder tiefer akzeptiert das WR-System nicht. Eine weitere Erkenntnis ist, dass die höchste Trefferquote auch erst bei Eingabe des T-Schlüssels erfolgt. Das bedeutet, dass er am aussagekräftigsten ist. Wird der Technologieschlüssel mit einem Schlagwort oder einem Beispieldokument kombiniert, bleibt die gefundene Anzahl der Known-Items gleich, die Anzahl der Gesamttreffer verringert sich jedoch deutlich.

Bei Eingabe eines Beispieldokuments ist die Trefferqualität bei einem aussagekräftigen Beispiel sehr gut, bei einem schlechter passenden dementsprechend niedriger. Auch die Übereinstimmung der Ergebnisse mit dem Eingabewert nimmt innerhalb der Trefferliste nach unten hin ab. Es kann davon ausgegangen werden, dass die unteren Treffer bei weitem nicht mehr so relevant sind wie die obersten Treffer. Bei einer Kombination von mehreren Beispieldokumenten wird deutlich, dass das Ergebnis umso besser wird, je mehr Beispiele hinzugenommen werden und je spezifischer diese für das T-Schlüsselprofil sind. Die Availability wird somit höher.

Bei der letzten Eingabeart mittels Schlagwörtern offenbart sich, dass die Trefferqualität ebenfalls besser wird, je genauer ein Schlagwort ist. Bei zu allgemeinen Anfragen fallen die Trefferzahlen generell sehr hoch und die Zahl der wirklich relevanten Patente ist sehr klein. Ist das Schlagwort jedoch ziemlich genau und aussagekräftig, kommt das Ergebnis sehr nahe an das Ergebnis einer reinen T-Schlüssel-Eingabe heran.

Zur Ansicht der einzelnen Ergebnistabellen wird auf Anhang A verwiesen.

6.2.3 Schwellwertbestimmung

Bisher ist nur die reine Funktionalität des What's-Related-Systems betrachtet worden. Ein weiterer Aspekt, der sich daraus ergibt, ist die Schwellwertbestimmung. Mit diesem Wert wird bestimmt, bis zu welchem Ähnlichkeitswert die Trefferergebnisse einer Suchanfrage beachtet werden.

Der derzeitige Schwellwert des Text-Mining-Verfahrens ist auf 10 Prozent gesetzt. Alle erhaltenen Ergebnisse, die zu der Suchanfrage eine Ähnlichkeit unterhalb der 10 Prozent aufweisen, werden nicht in die Ausgabemenge aufgenommen, sie werden sozusagen „abgeschnitten" und nicht angezeigt. Den Wert kleiner als 10 Prozent zu setzen, macht keinen allzu großen Sinn, da auf den unteren Treffermengen viel unnötiger Ballast - unpassend zur Suchanfrage - mitgeführt wird. Jedoch entspricht es den Anforderungen von IPM/C, diesen Wert so gering wie möglich zu halten, damit eine umfassende Analyse der Availability möglich ist.

Da dieser Punkt bereits abgearbeitet ist, stellt sich jetzt die Frage, bei welchem Schwellwert generell noch relevante Treffer durch WR gefunden und klassifiziert werden können.

Ferner kommt die Frage auf, wie das System für einen eventuellen Einsatz konfiguriert werden müsste. Gibt es einen allgemeingültigen Wert, der für alle Profile gilt oder muss der Wert für die unterschiedlichen Technologieschlüssel jeweils einzeln bestimmt werden?

Der Wert wird deshalb im weiteren Verlauf der Untersuchung für die drei Technologieschlüssel bestimmt und durch eigens erstellte Diagramme visuell dargestellt. Darüber hinaus sollen die Diagramme die Funktion übernehmen, auf einen Blick erkennen zu können, wo der besagte Schwellwert für das jeweilige Profil zu setzen ist. Im Anschluss an die Einzelergebnisse soll mittels einer Gesamttabelle eine Zusammenfassung der Erkenntnisse erfolgen.

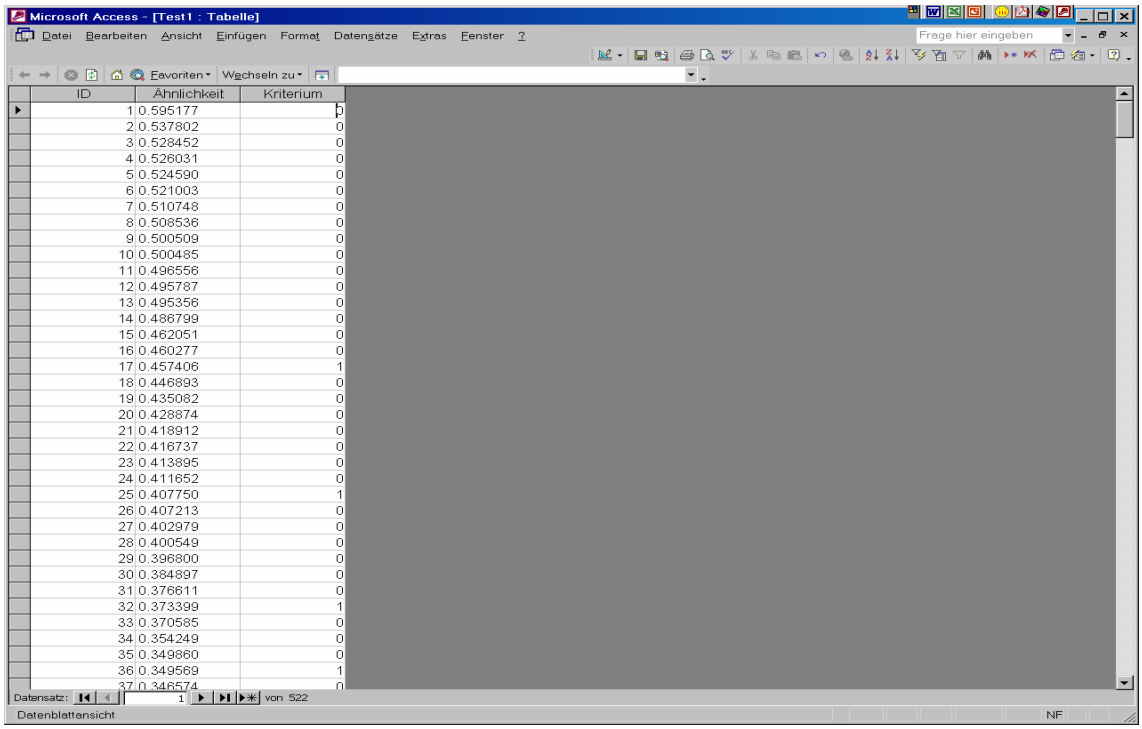
Vor der eigentlichen Durchführung werden allerdings noch einige Vorbereitungen für eine verbesserte Auswertung der Schwellwertergebnisse getroffen.

6.2.3.1 Vorbereitungen

Für diese zweite konstruktive Maßnahme, die im Rahmen der Evaluierung durchgeführt wird, ergibt sich die Notwendigkeit, MS Excel Diagramme zu erstellen, um eine eindeutige Interpretation der Schwellwertergebnisse zu ermöglichen.

Als Grundlage dient dazu die MS Access Datenbank, die bereits für die Auswertung der Known-Item-Analyse erstellt wurde.

Wenn abermals eine Suchanfrage an das WR-System gestellt wird und die Treffer mittels des Export-Links in die Datenbank überführt werden, erfolgt erneut der Abgleich auf übereinstimmende Datensätze der Trefferergebnisse und der Originaltabelle des bearbeiteten T-Schlüsselprofils (siehe dazu auch Kapitel 6.2.2.1). Der Unterschied bei der vorliegenden Verarbeitung liegt jedoch darin, dass nicht alle Merkmale ausgegeben werden sollen. Die nachfolgende Abbildung stellt das neue Ergebnis der Abfrage dar.



ID	Ähnlichkeit	Kriterium
1	0.595177	0
2	0.537802	0
3	0.528452	0
4	0.526031	0
5	0.524590	0
6	0.521003	0
7	0.510748	0
8	0.508536	0
9	0.500509	0
10	0.500485	0
11	0.496556	0
12	0.495787	0
13	0.495356	0
14	0.486799	0
15	0.462051	0
16	0.460277	0
17	0.457406	1
18	0.446893	0
19	0.435082	0
20	0.428874	0
21	0.418912	0
22	0.416737	0
23	0.413895	0
24	0.411652	0
25	0.407750	1
26	0.407213	0
27	0.402979	0
28	0.400549	0
29	0.396800	0
30	0.384897	0
31	0.376611	0
32	0.373399	1
33	0.370585	0
34	0.354249	0
35	0.349860	0
36	0.349569	1
37	0.346574	0

Abbildung 33: Übereinstimmende Datensätze der Treffer und der Originaltabelle

Dargestellt werden drei Spalten mit den Merkmalen „ID“, „Ähnlichkeit“ und „Kriterium“. Die ersten beiden Felder bezeichnen die Treffer-ID und den Grad der Ähnlichkeit eines Patentdokuments zu einer Suchanfrage. Diese Spalten sind auch in der ursprünglichen Abfrage aus Kapitel 6.2.2.1 enthalten. Mit der neuen Spalte „Kriterium“ wird nun gekennzeichnet, wo eine Übereinstimmung mit der Originaltabelle erfolgt und bei welchem Wert keine Übereinstimmung zu finden ist. Der Wert „0“ kennzeichnet dabei keine Übereinstimmung, und der Wert „1“ kennzeichnet eine bestehende Übereinstimmung.

Als letzter Schritt wird die Tabelle in eine MS Excel Arbeitsmappe überführt. Das Ziel dieser Vorbereitung liegt letztendlich darin, aus den übereinstimmenden Daten Diagramme zu erstellen, die die zu untersuchenden Schwellwerte visuell darstellen sollen.

6.2.3.2 Vorgehen bei der Schwellwertbestimmung

Um die Untersuchung an einem Beispiel aufzuzeigen, wird das Technologieschlüsselprofil Touchpad ausgewählt.

In einer Suchanfrage (siehe dazu Kapitel 6.2.2.2.2) werden mit Hilfe des Technologieschlüssels alle bereits klassifizierten Patentdokumente der Lernkollektion herausgefiltert. Innerhalb dieser Menge besteht die Möglichkeit, dass Teile davon falsch klassifiziert wurden - was letztendlich daran liegt, dass bei den irrelevanten Treffern *matching words* auftauchen, die denen der relevanten Treffer sehr ähnlich sind.

Nach Überarbeitung der erhaltenen Treffer nach Kapitel 6.2.3.1 wird ein Ergebnisdiagramm erstellt, das für die weitere Bearbeitung als Ausgangsgrundlage dient.

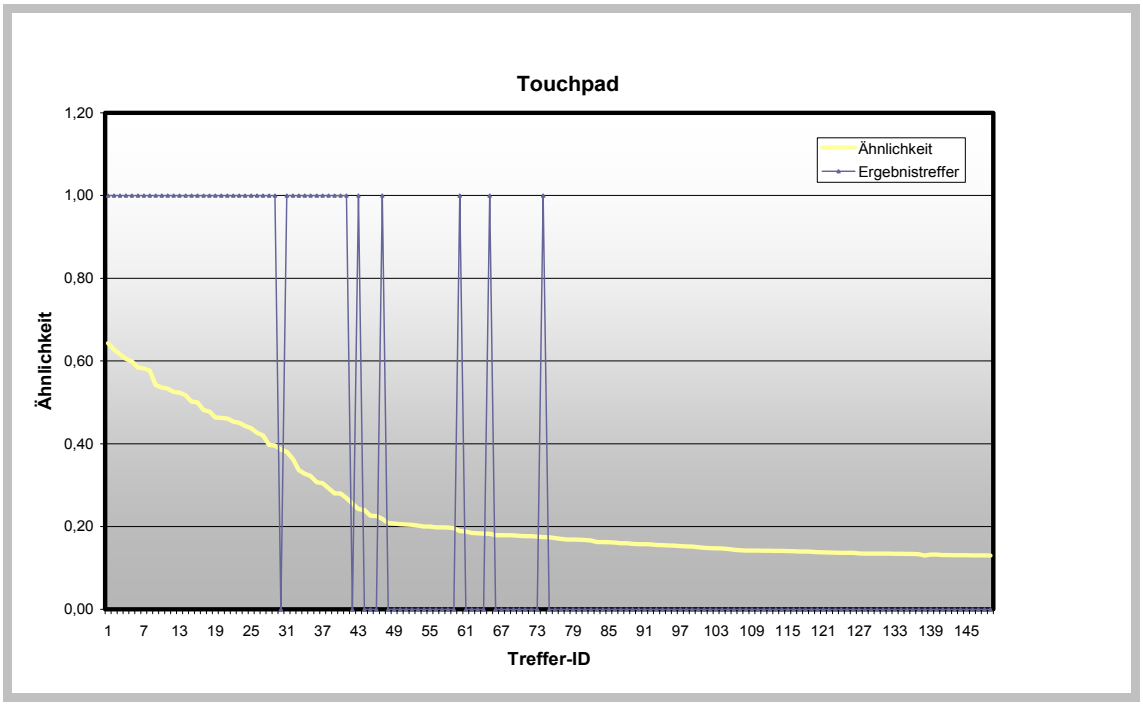


Abbildung 34: Recherche mit T-Schlüssel in der Lernkollektion bei $S_{Default} = 10\%$ [105]

❖ Was ist zu sehen?

Abbildung 34 stellt das Ergebnisdiagramm mit einem Schwellwert $S_{Default}$ von 10 Prozent dar. In dem Diagramm ist eine Linie in gelber Farbe ersichtlich, die den Verlauf der Ähnlichkeit aufzeigt. Sie nimmt dabei im weiteren Verlauf an Wert ab und pendelt sich gegen Ende ein. Des Weiteren ist mit der linken Rubrikenachse (y-Achse) der Wert für die Ähnlichkeit festgelegt. Auf der unteren Größenachse (x-Achse) sind die

[105] Quelle: Eigene Darstellung

Treffer mit ihrer Identitätsnummer oder Rankingnummer festgelegt. Diese Nummer gibt an, auf welchem Rang sich ein Treffer innerhalb der Gesamttrefferzahl befindet.

Weiterhin sieht man eine blaue Linie, die bei einer Ähnlichkeit von 1,00 ganz links zahlreiche Ausschläge aufweist, gefolgt von einem ersten Ausschlag nach unten auf die 0,00 bei Treffer-ID 30. Anschließend folgt ein zweiter größerer Block mit Ausschlägen auf die 1,00. Das „Auf und Ab“ passiert noch einige Male, bis schließlich ab Treffer-ID 75 keine Ausschläge mehr nach oben erfolgen; die Werte bleiben auf 0,00.

❖ Was wird daraus geschlossen?

Bei der blauen Linie handelt es sich um die Ergebnistreffer. Ein Ausschlag nach oben auf die 1,00 bedeutet ein richtig klassifiziertes Patentdokument, es gehört eindeutig zum Touchpad-Profil. Ein Ausschlag allerdings auf die 0,00 entspricht einem falsch klassifizierten Patentdokument und gehört nicht dem Touchpad-Profil an. Aus dem Diagramm ergibt sich letztendlich, dass alle 45 klassifizierten Patentdokumente der Lernkollektion richtig aufgefunden werden, aber auch 141 falsche Dokumente zu der Menge des Touchpad-Profils gezählt werden.

❖ Weiteres Vorgehen

In einem nächsten Schritt wird nun überprüft, welcher Ähnlichkeitswert W_1 bei dem ersten falschen Treffer der Lernkollektion auftritt und an welcher Stelle dieser Wert innerhalb einer Suchanfrage (nach Kapitel 6.2.2.2.1) in der Testkollektion auftaucht. Die Frage stellt sich, ob der Wert der Lernkollektion als Vorlage für den zu setzenden Schwellwert genutzt werden kann, denn offensichtlich ist, dass alle Treffer vor diesem ersten Fehltritt in der Lernmenge richtig klassifiziert sind. Im Weiteren soll nun überprüft werden, ob mit diesem Wert innerhalb der Testkollektion auch nur relevante Patentdokumente oder bereits Fehltreffer gefunden werden. Der Fehltreffer der Lernmenge wird somit als Vorlage für einen potentiellen Schwellwert benutzt.

Der erste Fehltreffer der Lernkollektion erscheint bei einem Wert von 39 Prozent. Innerhalb der Testkollektion werden bei diesem Ähnlichkeitswert folgende Patentdokumente getroffen.

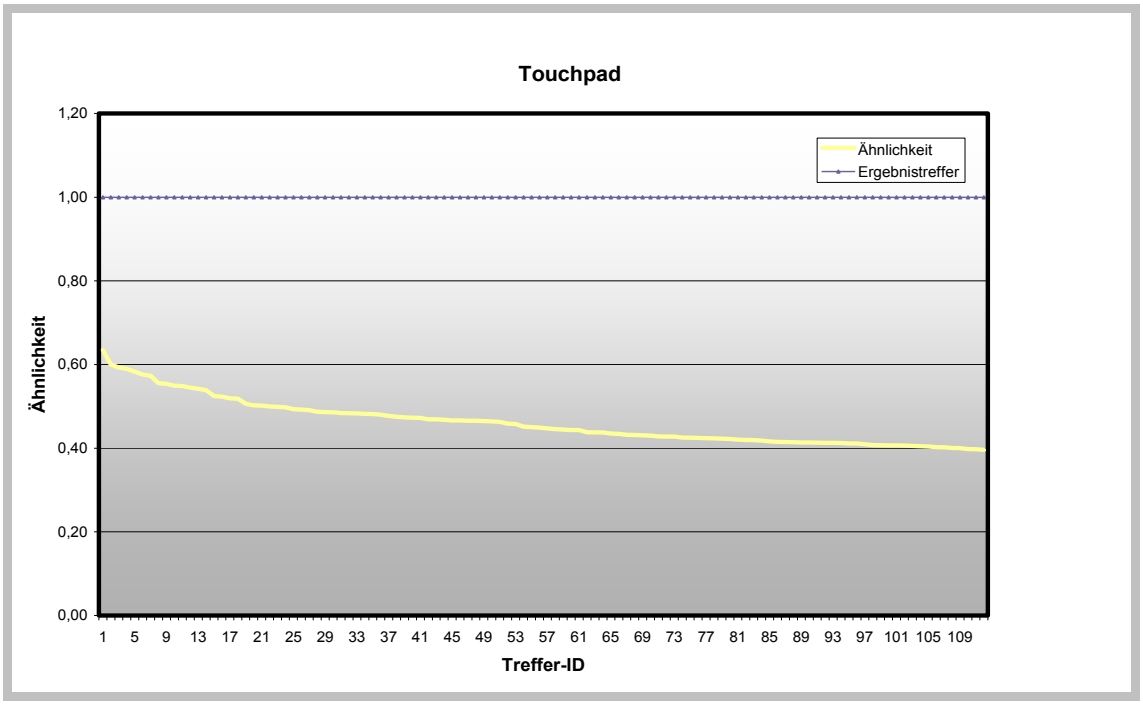


Abbildung 35: Recherche mit T-Schlüssel in der Testkollektion bei $W_1 = 39\%$ [106]

❖ Was ist zu sehen?

Abbildung 35 zeigt die Ergebnisse für einen Schwellwert von 39 Prozent innerhalb der nichtklassifizierten Patente an. Bei einem derartigen Wert tauchen lediglich richtig klassifizierte Patente auf. Sämtliche vorkommenden Ausschläge liegen bei 1,00.

❖ Was wird daraus geschlossen?

Damit zeigt sich, dass alle 112 getroffenen Patentdokumente aus der Testkollektion richtig klassifiziert werden können. Die 39 Prozent könnten nun als Endwert verwendet werden, jedoch ist offensichtlich, dass Potenzial für einen niedrigeren Wert vorhanden ist. Ein niedrigerer Wert wäre besser, da einem T-Schlüssel somit mehr Patentdokumente zugeordnet werden könnten.

❖ Weiteres Vorgehen

Deshalb wird in einem weiteren Durchgang nach demselben Verfahren der zweite Fehltreffer der Lernkollektion aus Abbildung 34 überprüft. Hierbei liegt der Ähnlichkeitswert W_2 bei 26 Prozent.

[106] Quelle: Eigene Darstellung

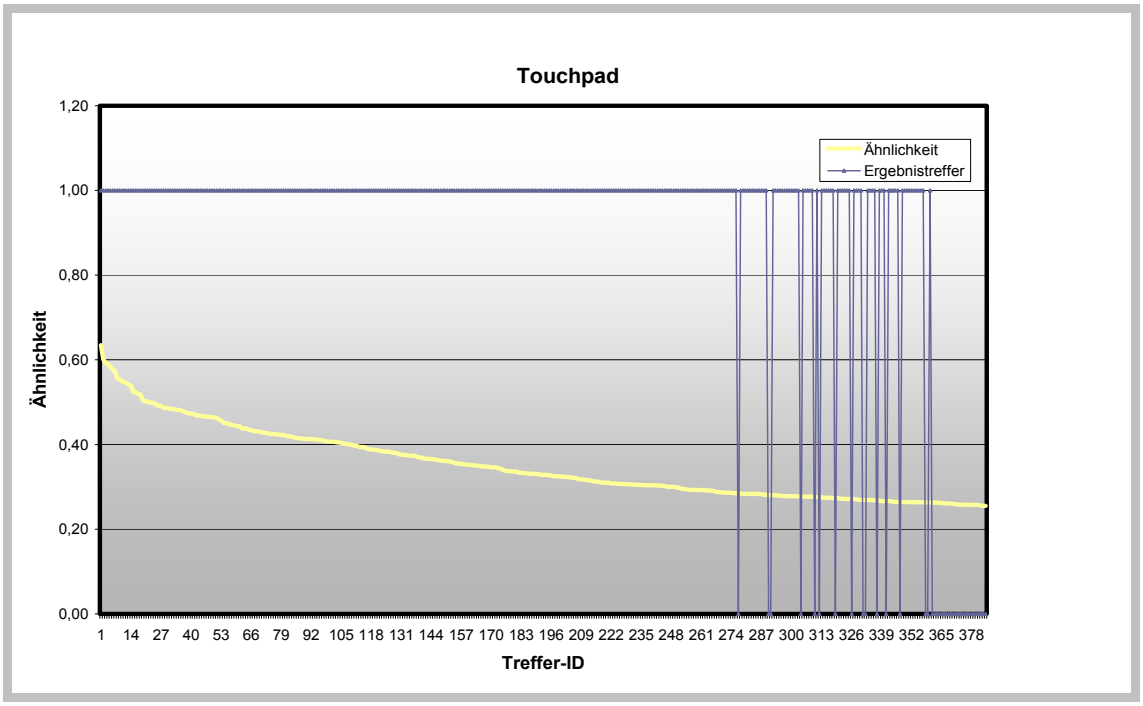


Abbildung 36: Recherche mit T-Schlüssel in der Testkollektion bei $W_2 = 26\%$ [107]

❖ Was ist zu sehen?

Das Ergebnis für einen Schwellwert von 26 Prozent sieht wie erhofft anders aus. Erkennbar ist auch, dass weit mehr Treffer gemacht werden als im vorigen Diagramm. Der erste Fehltreffer tritt in Diagramm 36 bei einem Ähnlichkeitswert von 28 Prozent bei Treffer-ID 277 auf. Danach erscheinen in einem zweiten Block abermals relevante Treffer. Die darauf folgenden Treffer sind erneut ungenau, und ein Wechsel zwischen relevant und irrelevant schließt sich an. Bis zum letzten Treffer mit Nummer 385 gibt es nur noch falsche Treffer. Innerhalb der Gesamtmenge werden damit 345 Patentdokumente richtig und 40 falsch klassifiziert.

❖ Was wird daraus geschlossen?

Für das gesuchte Ergebnis bedeutet das aber letztendlich, dass ein Schwellwert von 29 Prozent ein Optimum an richtig klassifizierten und relevanten Patenten herausbringt. Demnach muss der Schwellwert auf 29 Prozent gesetzt werden. Der überprüfte Wert von 26 Prozent kann nicht verwendet werden, da ansonsten falsch klassifizierte Patente übernommen würden.

Für das WR-System hat dieser neu bestimmte Endwert zu Folge, dass die gesamte Lern- und Testmenge des Touchpad-Profils nur bis zu diesem Wert verarbeitet oder beachtet werden darf. Von dem neuen Wert ausgehend, können dann schließlich insgesamt 275 nichtklassifizierte Dokumente aus der Testkollektion dem Touchpad-Profil zugeordnet werden. Treffer-ID 276 ist nach genauerer Betrachtung zwar auch ein relevanter Treffer, besitzt aber einen Ähnlichkeitswert von 28 Prozent. Da aber alle

[107] Quelle: Eigene Darstellung

restlichen Patentdokumente unterhalb der 29 Prozent unberücksichtigt bleiben, fällt der Treffer weg. Eine weitere Erkenntnis kann auch aus dem Ergebnis gezogen werden: Die ursprünglich bekannte Menge von 12 relevanten Profilen in der Testkollektion zum Thema Touchpad wurde bei weitem übertroffen.

Im Rahmen einer Untersuchung der Patente unterhalb des festgelegten Schwellwerts ergibt sich eine weitere interessante Beobachtung. Die Treffer, die eigentlich nach unten auf die Null ausschlagen, sind eindeutig dem Oberbegriff bzw. in der übergeordneten Ebene des Technologieschlüssels anzusiedeln. Sie sind demnach nicht unbedingt falsch. Diesem Phänomen sollte deshalb im nächsten Schritt der Evaluierungsphase Beachtung geschenkt werden bzw. der Aspekt sollte genauer überprüft werden. Für die vorliegende Arbeit kann der Beobachtung nicht weiter nachgegangen werden, da die Datenmenge des WR-Systems zuvor erweitert werden müsste.

6.2.3.3 Weitere Ergebnisse

Auf Grund der sich wiederholenden Vorgänge zur Festlegung des Schwellwerts soll der Ablauf der nachfolgenden Untersuchungen nicht ein weiteres Mal beschrieben werden. Die Ergebnisse werden lediglich in Form der Abschlussdiagramme dargestellt und erläutert.

6.2.3.3.1 Suchanfrage bei Eingabe mehrerer Beispieldokumente

Als Recherchekriterium werden fünf Beispieldokumente verwendet, die spezifisch für den Technologieschlüssel Touchpad sind.

Bei dieser Vorgehensweise kann jedoch mit Hilfe des ersten und zweiten Fehltreffers kein Schwellwert bestimmt werden, da erneut nur relevante Treffer auftauchen. Deshalb wird ein dritter Fehltreffer mit einem Ähnlichkeitswert W_3 von 23 Prozent dazu genommen.

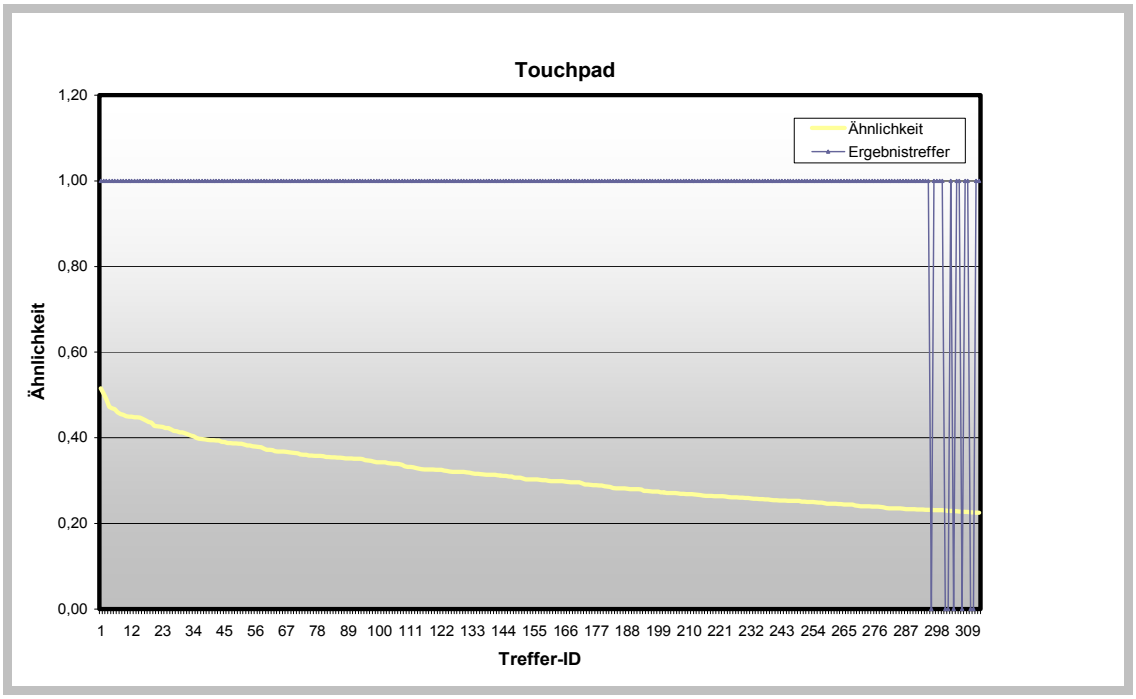


Abbildung 37: Recherche mit Dokumenten in der Testkollektion bei $W_3 = 23\%$ [108]

❖ Was ist zu sehen?

In Abbildung 37 wird die Testkollektion mit einem gesetzten Schwellwert von 23 Prozent überprüft. Dabei ist ein erster großer Block mit richtig zugeordneten Patentdokumenten zu sehen. Ab Treffer-ID 296 erscheinen jedoch die falsch klassifizierten Dokumente. Anschließend folgt ein Wechsel zwischen relevanten und nichtrelevanten Patenten. Insgesamt werden mit dem Wert 306 Patentdokumente richtig und 8 falsch erkannt. Somit muss überprüft werden, welcher Ähnlichkeitswert bei Treffer-ID 296 auftritt und daraus der Schwellwert bestimmt werden.

❖ Was wird daraus geschlossen?

Fehltreffer Nummer 296 hat einen Ähnlichkeitswert von 23 Prozent. Deshalb wird für die Eingabe mittels fünf Beispieldokumenten ein Schwellwert von 24 Prozent festgelegt. Mit diesem Optimum ist es möglich, 284 der getroffenen 314 Patentdokumente für das Touchpad-Profil zu klassifizieren. Die 11 Treffer zwischen dem letzten Dokument, das zu den 24 Prozent gehört, und dem ersten Fehldokument mit den 23 Prozent werden fallen gelassen, da sie auch einen Ähnlichkeitswert von 23 Prozent besitzen.

Des Weiteren werden auch hier wieder eindeutig mehr mögliche Klassifikationen aufgedeckt als die 12 bekannten Profile der Testkollektion.

[108] Quelle: Eigene Darstellung

6.2.3.3.2 Suchanfrage bei Eingabe des höchsten Precision-Werts

Bei der nachfolgenden Untersuchung soll auf die Precision als weiterem wichtigem Aspekt eingegangen werden. Eine hohe Precision weist generell darauf hin, dass die gefundenen und relevanten Treffer einer Suchanfrage in einem ausgewogenen Verhältnis zur Gesamttrefferzahl stehen.

Dazu wird in den Ergebnistabellen der Known-Item-Analyse (siehe dazu Anhang A) der höchste Wert herausgefiltert. Dieser Wert ergibt sich durch Kombination der Schlagwörter „input instruction“ und „touch“. Die beiden Begriffe werden deshalb als Rechercheкriterium verwendet. Bei dieser Variante kann mit Hilfe des ersten Fehltreffers der Schwellwert bestimmt werden.

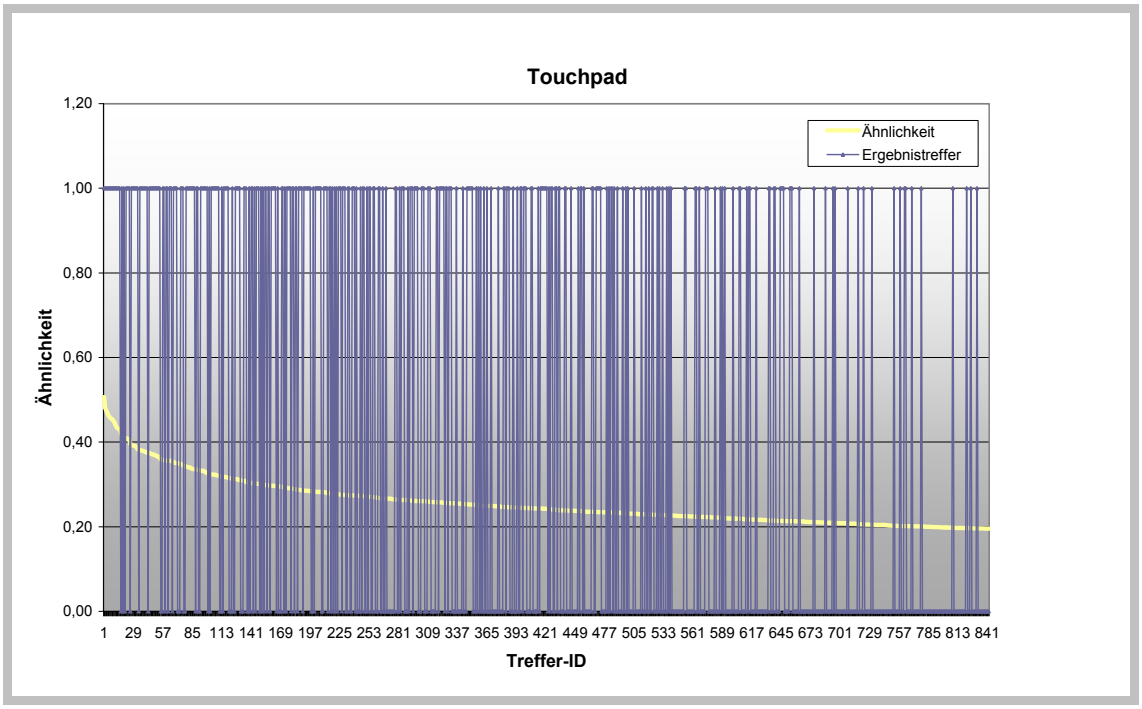


Abbildung 38: Recherche mit Begriffen in der Testkollektion bei $W_1 = 20\%$ [109]

❖ Was ist zu sehen?

Die Testkollektion in Diagramm 38 wird mit einem gesetzten Schwellwert von 20 Prozent untersucht, und es besteht offensichtlich das erste Mal ein anderes Schema als bei den anderen Diagrammen. Zu Beginn der Treffermenge besteht ein relativ kleiner Block mit 16 richtigen Treffern, Nummer 17 ist bereits ein Fehltreffer. Danach folgt nur noch ein Wechsel zwischen 241 richtig und 596 falsch klassifizierten Patentdokumenten. Somit muss überprüft werden, welcher Ähnlichkeitswert bei Treffer-ID 17 auftritt und daraus der Schwellwert bestimmt werden. Darüber hinaus wird auch eine relativ hohe Trefferzahl mit insgesamt 843 Dokumenten festgestellt.

[109] Quelle: Eigene Darstellung

- Was wird daraus geschlossen?

Durch das Diagramm wird für den Schwellwert ein Maß von 43 Prozent ermittelt. Mit diesem Optimum ist es möglich, insgesamt 16 Patentdokumente für das Touchpad-Profil zu klassifizieren. Treffer-ID 17 weist schon einen Ähnlichkeitswert von 42 Prozent auf. Damit können jedoch vier Dokumente zusätzlich klassifiziert werden, da auch hier bereits bekannt war, dass 12 Treffer auf jeden Fall gefunden werden müssen.

6.2.3.4 Zusammenfassung

Lernkollektion				
Schwellwert	**Known Items**	**Richtige Treffer**	**Falsche Treffer**	**Prozent**
Eingabe des T-Schlüssels				
10%	45	45	141	100
1. Fehltreffer: 39%	45	29	1	64,4
2. Fehltreffer: 26%	45	40	2	88,9
Eingabe von fünf Beispieldokumenten				
10%	45	43	201	95,6
1. Fehltreffer: 33%	45	20	1	44,4
2. Fehltreffer: 25%	45	32	2	71,1
3. Fehltreffer: 23%	45	32	3	71,1
Eingabe des höchsten Precision-Werts				
10%	45	42	76	93,3
1. Fehltreffer: 20%	45	20	1	44,4
Testkollektion				
Eingabe des T-Schlüssels				
10%	Menge unbekannt, mindestens 12	395	4.456	100
39%	Menge unbekannt, mindestens 12	112	0	28,4
26%	Menge unbekannt, mindestens 12	345	40	87,34
Ergebnis: Schwellwert von 29% ergibt 275 neue Klassifizierungen				
Eingabe von fünf Beispieldokumenten				
10%	Menge unbekannt, mindestens 12	397	2.104	100
33%	Menge unbekannt, mindestens 12	122	0	30,7
25%	Menge unbekannt, mindestens 12	264	0	66,5
23%	Menge unbekannt, mindestens 12	306	8	77,1
Ergebnis: Schwellwert von 24% ergibt 284 neue Klassifizierungen				
Eingabe des höchsten Precision-Werts				
10%	Menge unbekannt, mindestens 12	376	1.648	100
20%	Menge unbekannt, mindestens 12	247	596	41,4
Ergebnis: Schwellwert von 43% ergibt 16 neue Klassifizierungen				

Tabelle 7: Zusammenfassung der Ergebnisse für das Touchpad-Profil[110]

[110] Quelle: Eigene Darstellung

In Tabelle 7 sind die Ergebnisse der Schwellwertbestimmung für das Touchpad-Profil zusammengefasst. Zuerst werden die Untersuchungen für die Lernkollektion bei Eingabe des T-Schlüssels, der fünf Beispieldokumente und bei Eingabe der Begriffe mit der höchsten Precision dargestellt. Des Weiteren wird in dem Zusammenhang angegeben, wie viele Known-Items innerhalb der beiden Kollektionen vorkommen und wie viele davon bei einem bestimmten Schwellwert getroffen werden. Das entspricht dem Feld „Richtige Treffer“. Ein falscher Treffer wird vermerkt, wenn innerhalb der Treffermenge keine Known-Items vorkommen. Das letzte Feld entspricht dem Prozentsatz, mit dem die richtigen Known-Items erkannt werden. Als Erstes werden die Ergebnisse für den momentanen Schwellwert von 10 Prozent aufgeführt. Anschließend erfolgt die Überprüfung der ersten Fehltreffer. Die Schwellwerte der Fehltreffer werden im Weiteren für die Testkollektion überprüft. Für diese Kollektion ist dabei nur die Mindestmenge der Known-Items bekannt, nicht aber, wie viele insgesamt innerhalb der Testmenge vorkommen. Auf Grund dessen kann für den derzeitigen Schwellwert von 10 Prozent nur eine ungefähre Angabe gemacht werden. Aus den gewonnenen Erkenntnissen der Testkollektion kann dann letztendlich der Schwellwert bestimmt werden.

6.2.3.5 Besonderheit

Dieselben Untersuchungen werden entsprechend für die beiden anderen Technologieschlüssel Laserschweißen und Fußgängerschutz ausgeführt. Für die Ergebnisse der Schwellwertbestimmung des dritten Profils Fußgängerschutz wird allerdings auf Anhang B verwiesen.

Eine Besonderheit, auf die im folgenden Unterkapitel eingegangen wird, zeigt sich beim Laserschweißen-Profil.

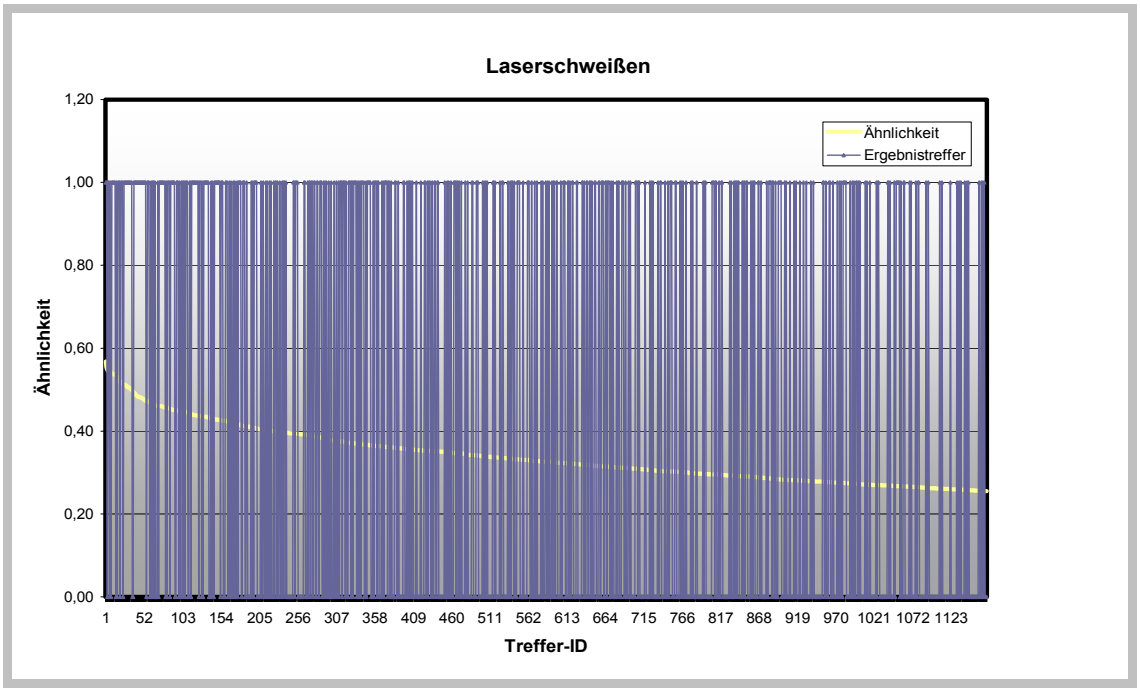


Abbildung 39: Recherche mit T-Schlüssel in der Testkollektion bei $W_1 = 32\%$ [111]

❖ Was ist zu sehen?

Die Testkollektion in Diagramm 39 weist mit einem gesetzten Schwellwert von 32 Prozent ein ähnliches Schema auf wie die Suchanfrage mittels des höchsten Precision-Werts für das Touchpad-Profil. Der Unterschied liegt darin, dass nur ein wechselndes „Auf und Ab“ der Treffer zu sehen ist. Des Weiteren ist eine hohe Trefferzahl mit insgesamt 1.170 Patentdokumenten aufgetreten, wovon 478 der insgesamt 894 bereits bekannten Laserschweißen-Dokumente richtig erkannt werden. Für die Ergebnismenge des Diagramms bedeutet das dann folglich, dass 692 Treffer falsch erkannt werden. Es ist aber kein größerer zusammenhängender Block an richtigen oder falschen Treffern ersichtlich.

[111] Quelle: Eigene Darstellung

- ❖ Was wird daraus geschlossen?

Das Diagramm weist augenscheinlich nicht dieselben Merkmale wie die vorherigen Ergebnisse auf. In diesem Fall lässt sich kein Schwellwert finden, auch wenn die Daten bis in höhere Treffermengen verfolgt werden.

6.2.4 Einfluss der Internationalen Patentklassifikation

Der Einfluss der IPC innerhalb des Patentwesens ist, wie bereits in Kapitel 3.1.3 aufgezeigt, sehr groß. Durch die Änderungen der Internationalen Patentklassifikation ist ein Auf- und Nachbearbeiten der Rechercheprofile innerhalb von IPM/C in regelmäßigen Abständen nötig. Das ist zum einen kostenintensiv, zum anderen auch sehr zeitaufwendig.

Ab 01. Januar 2005 sollte eine neue reformierte Ausgabe der Internationalen Patentklassifikation in Kraft treten, der Termin wurde jedoch auf den 01. Januar 2006 verschoben. Ab dann werden die Anpassungen nicht wie bisher alle fünf Jahre vorgenommen, sondern für 2006 in jedem Quartal, ab 2007 dann in jedem Monat.

Den Aufwand, der sich aus dieser schwerwiegenden Entscheidung ergibt, kann IPM/C allerdings nicht mehr betreiben.

Auch im WR-System ist die IPC Bestandteil, da aus ihr Textmerkmale der Patentdokumente extrahiert werden. So kam die Frage auf, ob das Text-Mining-Verfahren eventuell auch ohne Beachtung der Internationalen Patentklassifikation ansprechende Ergebnisse liefern kann. Um das zu überprüfen, wird die Lernmenge ein zweites Mal in das WR-System integriert, mit dem Unterschied, dass alle Merkmale, in der die IPC enthalten ist, nicht beachtet und weggelassen werden.

Diese zweite Testversion ist deshalb über eine weitere Weboberfläche zu erreichen.

Untersucht wird in den folgenden Unterkapiteln, inwiefern sich das Verfahren im Gegensatz zu der ersten Testversion anders verhält. Weiterhin wird überprüft, ob die IPC zu einem besseren Ergebnis beiträgt oder ob sie entbehrlich ist.

6.2.4.1 Vorgehen bei der Untersuchung

Um den Einfluss genauer zu untersuchen, wird nach Kapitel 6.2.2.2.2 mit einer Recherche innerhalb der Lernkollektion vorgegangen. Als Eingabekriterium wird der Technologieschlüssel ausgewählt. Für den Testablauf selber wird das Fußgängerschutz-Profil verwendet. Der Vorgang wird zweimal ausgeführt – einmal für die bisher verwendete Testversion und das andere Mal für die neue Webversion, ohne Berücksichtigung der Merkmale, in der IPC-Klassen enthalten sind. Untersucht werden die Ergebnisse mit einem Schwellwert von 10 Prozent, damit so viele Treffer wie möglich gemacht werden können.

Auf die Ergebnisse des Laserschweißen- und Touchpad-Profils wird in Anhang C näher eingegangen.

6.2.4.2 Ergebnis

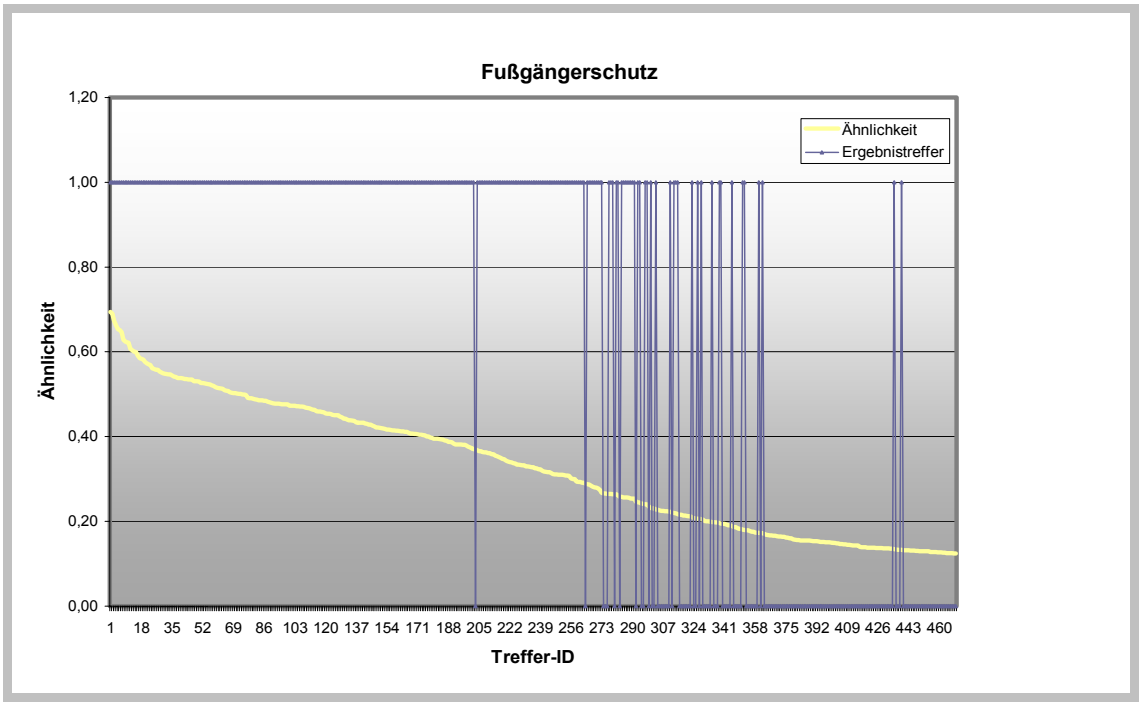


Abbildung 40: Recherche in der Lernkollektion bei $S_{Default} = 10\%$ mit IPC[112]

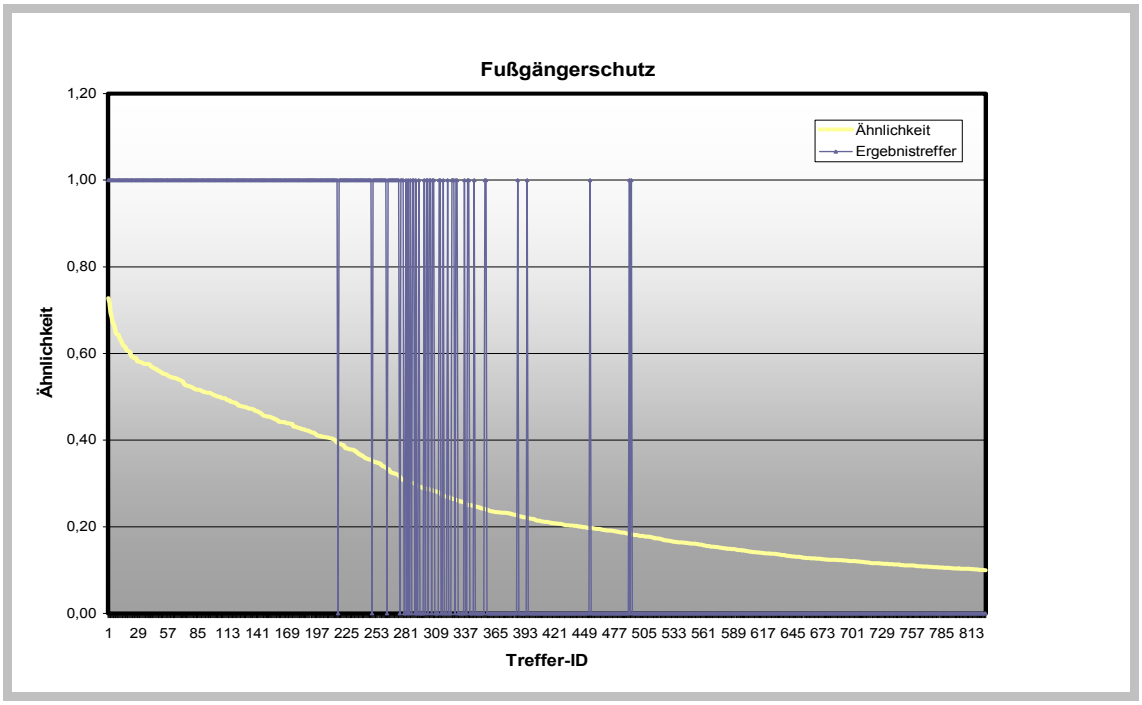


Abbildung 41: Recherche in der Lernkollektion bei $S_{Default} = 10\%$ ohne IPC[113]

[112] Quelle: Eigene Darstellung
[113] Quelle: Eigene Darstellung

❖ Was ist zu sehen?

Beide Abbildungen weisen überraschenderweise kaum große Abweichungen auf. Allerdings wird aus Abbildung 41 ersichtlich, dass die Trefferzahl mit insgesamt 826 Treffern doppelt so hoch ist. Ein falscher Ausschlag erscheint zum ersten Mal bei Treffer-ID 217, die Patentdokumente davor sind alle richtig erkannt worden. Insgesamt sind bei der Kollektion 308 richtige und 518 falsche Klassifikationen erfolgt. In Abbildung 40 erscheint der erste Fehltreffer bei einer Gesamtzahl von 469 Treffern bei Nummer 203. Hier sind 307 Patentdokumente richtig und 162 falsch erkannt worden. Sämtliche richtige Treffer unterscheiden sich somit erstaunlicherweise nur um ein Dokument. Genauso erstaunlich ist, dass sich die ersten Fehlstellen nur um 14 Patentdokumente unterscheiden. Des Weiteren sieht auch die Struktur sehr ähnlich aus. Zu Beginn erscheint ein großer, richtig klassifizierter Block, dann einige Fehltreffer - diese werden in der Folge zahlreicher, dann gibt es gegen Ende nur vereinzelt richtige Ausschläge und schließlich erscheinen nur noch die falschen Treffer.

❖ Was wird daraus geschlossen?

An der Stelle ist ein deutlicherer Unterschied zwischen den beiden Ergebnissen erwartet worden. Darüber hinaus zeigt sich, dass die Trefferstellen ziemlich nahe beieinander stehen. Auf dieses Phänomen soll deshalb im nächsten Kapitel ein genauerer Blick geworfen werden.

6.2.4.3 Trefferstellen

Im Rahmen des Tests zur Beeinflussung der IPC ergibt sich eine letzte Untersuchungsmöglichkeit. Dadurch soll geklärt werden, an welcher Stelle die jeweiligen Treffer der beiden Weboberflächen aufgefunden werden.

Vermutet wird, dass die weiter hinten stehenden Treffer aus der Ergebnisliste bei einer Recherche mit Berücksichtigung der IPC, eher an den vorderen Stellen in der Trefferliste bei einer Recherche ohne Berücksichtigung der IPC, aufzufinden sind. Der Grund liegt darin, dass bei der Recherche unter Einbeziehung der Internationalen Patentklassifikation die IPC-Merkmale eine höhere Gewichtung haben und deren *matching words* an den vordersten Stellen auftauchen. Bei den Treffern, die von der Relevanz weiter unten stehen, bekommen dafür die reinen Schlagwörter eine höhere Bedeutung, und die IPC-Klasse steht nicht mehr dabei.

Andererseits werden die hinteren Rangtreffer bei der Recherche ohne IPC, innerhalb der Recherche mit Berücksichtigung der IPC, weiter vorne aufgefunden. Der Grund liegt auch hier darin, dass die ersten Treffer, die keinerlei Merkmale der IPC aufweisen, reine Begriffe sind. Und diese Begriffe tauchen in der Recherche mit IPC an den hinteren Plätzen auf.

Um das im Weiteren zu klären, wird eine Recherche nach Kapitel 6.2.2.2.2 innerhalb der Lernkollektion ausgeführt. Suchkriterium ist der T-Schlüssel des Fußgängerschutz-Profils. Die Recherche wird dabei zunächst für die erste Testversion mit Beachtung der

IPC und im Anschluss für die zweite Testversion ohne Beachtung der IPC durchgeführt.

Aus den Ergebnissen werden insgesamt acht Beispieldokumente herausgefiltert und überprüft, an welcher Stelle sie in beiden Trefferlisten jeweils stehen. Das Endergebnis ist in der nachfolgenden Tabelle dargestellt.

Dokument	Mit IPC	Ohne IPC
1998-525073	265	256
2002-100423	273	236
1999-218518	345	386
2003-630670	439	454
2002-141085	240	251
2002-107190	303	313
1998-415329	362	395
1999-127201	351	356

Tabelle 8: Trefferstellen für das Fußgängerschutz-Profil[114]

Das Ergebnis daraus ist, dass die Treffer ziemlich nahe beieinander stehen und sich die vorher geäußerte Vermutung nicht bestätigt hat: Hintere Treffer stehen bei der anderen Klassifikationsmöglichkeit nicht weiter vorne, und vordere Treffer stehen auch nicht weiter hinten. Die Ergebnisse sind sich sehr ähnlich. Eine weitere daraus gewonnene Erkenntnis ist, dass die Treffer ohne Berücksichtigung der IPC ein bisschen weiter hinten stehen.

[114] Quelle: Eigene Darstellung

Das soll in einem Diagramm besser veranschaulicht werden:

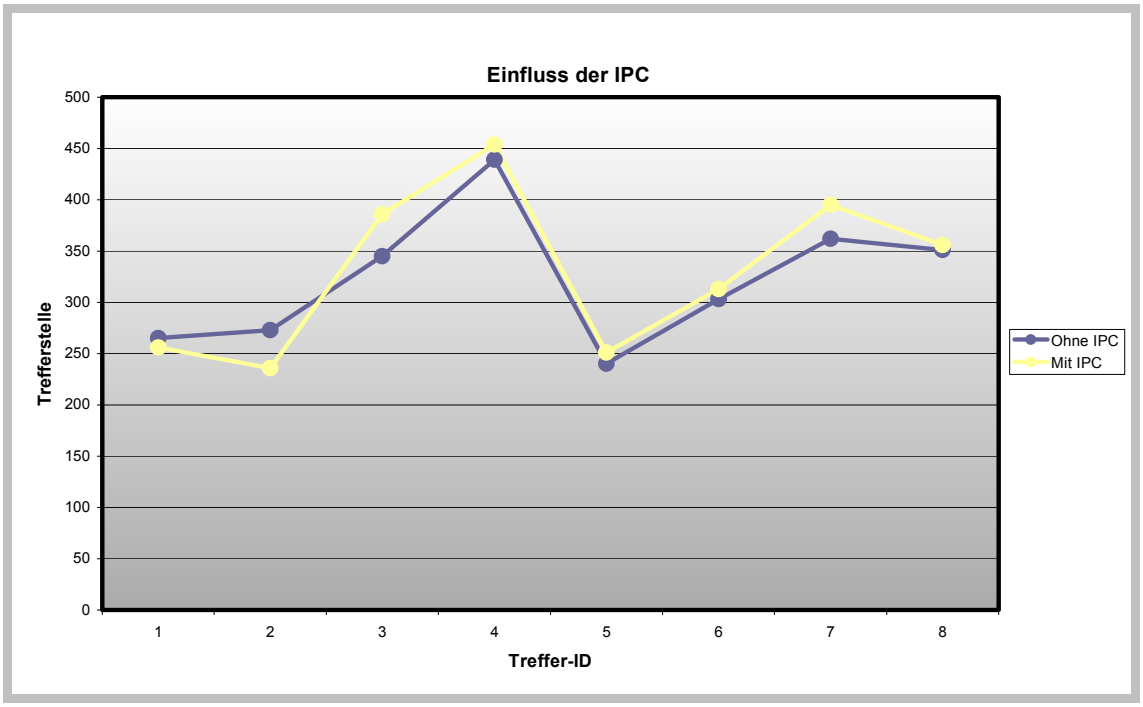


Abbildung 42: Vergleich der Trefferstellen für das Fußgängerschutz-Profil mit und ohne IPC-Berücksichtigung[115]

In Abbildung 42 ist zum einen eine Linie in gelber Farbe ersichtlich, die alle Treffer mit Berücksichtigung der IPC aufzeigt und zum anderen eine blaue Linie, die alle Treffer ohne Berücksichtigung der IPC darstellt. Des Weiteren sind auf der x-Achse die Treffer mit ihrer Identitätsnummer oder Rankingnummer und auf der y-Achse die genauen Trefferstellen aufgetragen. Beide Linien sind dabei fast identisch, die Abweichung ist minimal. Nur zu Beginn gehen beide Linien etwas weiter auseinander.

[115] Quelle: Eigene Darstellung

7 Diskussion

In diesem Kapitel werden die Ergebnisse der Evaluierungsvorgänge ausgewertet und hinsichtlich ihrer Effizienz bewertet. Darauf aufbauend wird auf Grund der gewonnenen Erkenntnisse eine Empfehlung ausgesprochen. Im Anschluss daran erfolgt ein kurzer Ausblick auf die weitere Nutzungsmöglichkeit.

7.1 Auswertung

Auf Grund der Problemsituation, die innerhalb der Patentabteilung besteht, wird eine mögliche Lösungsalternative in Form eines Text-Mining-Verfahrens ausgesucht. Es erfolgt eine Untersuchung des Systems, um festzustellen, ob durch den Einsatz des Verfahrens eine Verbesserung der Situation möglich ist.

Die erste dazu durchgeführte Maßnahme ist die Known-Item-Analyse. Diese Analyse hat das Ziel, die generelle Klassifikations- und Recherchefunktionalität des Verfahrens zu testen. Mit Hilfe der Trefferquote, der Availability, kann dazu ein aussagekräftiger Wert bestimmt werden. Er sagt aus, wie viel Prozent der relevanten Patentdokumente innerhalb der gesamten Trefferzahl wirklich getroffen werden.

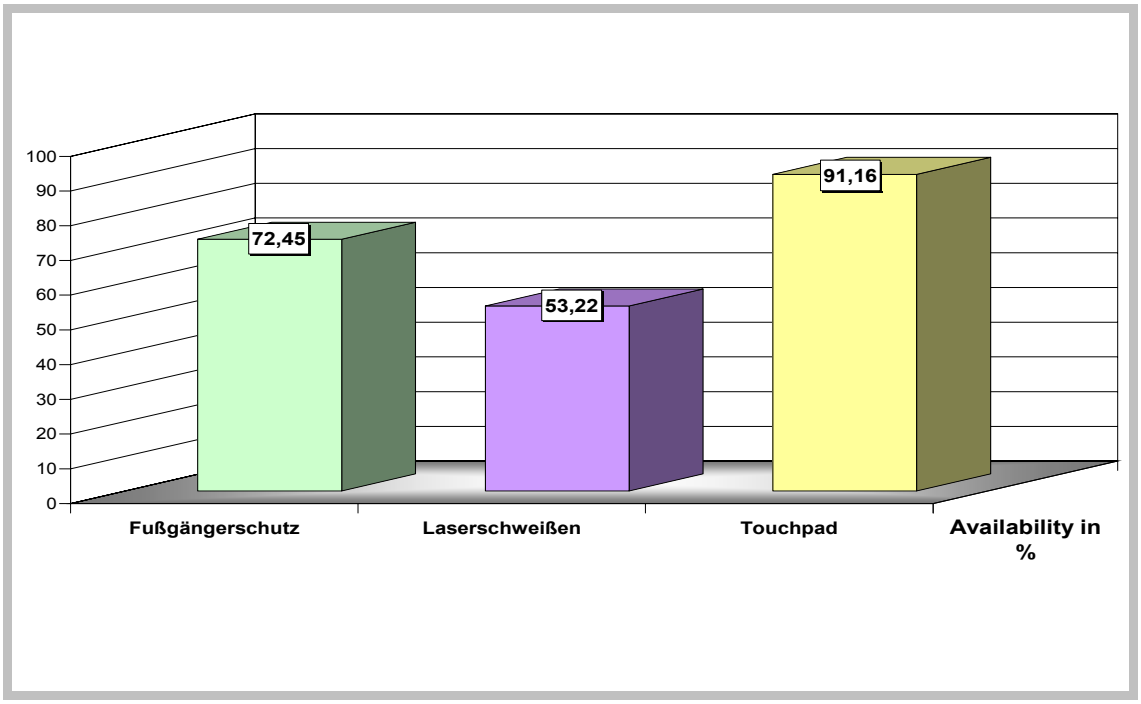


Abbildung 43: Endergebnis der Known-Item-Analyse[116]

Für die drei Technologieschlüssel-Profile ergeben drei unterschiedliche Werte.

[116] Quelle: Eigene Darstellung

Das Touchpad-Profil erzielt dabei den besten Wert mit insgesamt 91,16 Prozent. Dieses Ergebnis lässt sich damit erklären, dass das Profil im Verhältnis zu den anderen beiden Profilen eine relativ kleine Lernmenge besitzt. Mit insgesamt 45 Patenten ist das die geringste Menge. Das What's-Related-System scheint damit am besten zurechtzukommen, und nutzt alle daraus gewonnenen *matching words* zu einer Neuklassifizierung oder zu einer Auflistung der bereits klassifizierten Dokumente. Zwar werden bei der Recherche relativ viele Gesamttreffer erzielt, die relevanten Dokumente werden jedoch sehr oft bzw. fast immer gefunden. Interessant ist, dass die Ergebnisse immer besser werden, je aussagekräftiger und präziser ein Suchbegriff ist. Die Treffergesamtzahl einer Suchanfrage wird zum einen kleiner und aussagekräftiger, und zum anderen werden die Known-Items häufiger getroffen. Die Treffer werden jedoch schlechter, je unpassender ein Suchbegriff ausgewählt wird. Ist er zu allgemein gehalten, werden viele irrelevante Dokumente aufgefunden, und im Endeffekt wird so viel Ballast mitgenommen. Dies passiert deshalb, weil weit mehr Themengebiete getroffen als gesucht werden. Z.B. wird die Suchanfrage „passenger" eingegeben. Das Ziel mit Eingabe dieses Begriffes besteht darin, viele relevante Patentdokumente zum Thema Fußgängerschutz zu erhalten. Das Ergebnis mittels dieses Schlagwortes hat zur Folge, dass eine sehr hohe Gesamttrefferzahl erzielt wird, von den relevanten Dokumenten werden jedoch nur wenige gefunden - im Gegensatz zu einer Eingabe mit anderen weit spezifischeren Schlagwörtern.
Die besten Ergebnisse, also die höchste Trefferquote der Known-Items, werden erzielt, wenn für den Suchvorgang der T-Schlüssel eingegeben wird.

Das Fußgängerschutz-Profil weist mit 72,45 Prozent schon eine schlechtere Verfügbarkeit auf. Auch das lässt sich an der Lernkollektion erklären. Mit den 309 Dokumenten dieser Menge steht es genau zwischen den beiden anderen T-Schlüsselprofilen. Beim WR-System scheint es demnach problematisch zu sein, eine größere Anzahl an bekannten Patentdokumenten zu finden.

Das Laserschweißen-Profil schließlich erweist sich als schlechtestes Beispiel. Mit nur 53,22 Prozent hat es lediglich knapp mehr als die Hälfte aller relevanten Patente getroffen. Wenn man erneut die Lernmenge betrachtet, lässt sich das Ergebnis schnell erklären. Mit insgesamt 4.998 Beispielen handelt es sich um die größte Anzahl. Daraus zeigt sich aber, dass die *matching words*, die als Vergleichswert zu einer Suchanfrage verwendet werden, nicht alle verarbeitet werden können. Die Wörter, die allzu oft vorkommen, sind nicht mehr relevant und werden nicht weiter beachtet, genauso wie die Wörter, die zu selten vorkommen. Diese Begriffe fungieren dann als Stoppwörter. Somit werden bei den eigentlich zu treffenden Patentdokumenten nicht alle getroffen.

Letzten Endes ergibt sich daraus, dass die Availability besser ausfällt, wenn nicht allzu viele Patente als Lernmenge dienen. Allerdings darf die Menge auch nicht zu klein ausfallen, da die Klassifikation hierbei auch zu schlecht wird. Das System hat sich in dieser Hinsicht deshalb als akzeptabel erwiesen, da ein Availability-Wert von 72,28 Prozent herauskommt. Erhofft wurde jedoch ein besserer Wert.

Der Wert besagt des Weiteren, dass das Text-Mining-Verfahren nicht besser als der derzeitige Realisierungsprozess innerhalb von IPM/C ist, sondern eine annähernd gleiche Qualität aufweist. Der erhoffte Beweis bzw. die Erfüllung der Anforderung, dass die 10 bis 20 Prozent der nicht klassifizierbaren Patentdokumente auch klassifiziert werden können, kann jedoch nicht erbracht werden.

Allerdings wird das Durchsuchen der Informationsergebnisse deutlich erleichtert. Gerade unerfahrene Nutzer oder Personen, die mit einem bestimmten Fachgebiet und seiner Terminologie nicht vertraut sind, werden durch die einfach gehaltenen Eingabemöglichkeiten besser unterstützt und können auch ohne Vorkenntnisse der Retrieval-Sprache erfolgreich arbeiten. Auch dass die Suchergebnisse entsprechend gruppiert und angezeigt werden können, gewährleistet wiederum einen erheblich besseren inhaltlichen, aber auch schnellen Überblick.

Der zweite Untersuchungsgegenstand hingegen greift die Frage auf, wie das System am Optimalsten konfiguriert werden sollte. Als Maß dafür wird der Schwellwert bestimmt. Die Suchanfrage selber wird mittels Eingabe des jeweiligen T-Schlüssels, mit Eingabe von fünf Beispieldokumenten und mit Eingabe der Begriffe, die die höchste Precision aufweisen, ausgeführt.

Bei Eingabe des T-Schlüssels kann für das Touchpad- und das Fußgängerschutzprofil jeweils ein Schwellwert von 29 Prozent festgelegt werden. Das bedeutet, mit einem eingestellten Schwellwert von 29 Prozent lassen sich bei dieser Eingabevariante die meisten Patentdokumente klassifizieren. Für das Touchpad-Profil ergeben sich damit 275 Neuklassifizierungen, für das Fußgängerschutz-Profil lassen sich mit dem eingestellten Schwellwert 37 Neuklassifikationen vornehmen. Ein Wert unter den 29 Prozent hätte bei beiden Profilen zur Folge, dass zwar ein höherer Satz an Dokumenten klassifiziert werden könnte, aber auch mit falsch zugeordneten Patenten gerechnet werden müsste. Den Wert höher zu setzen, bringt zwar auch nur relevante Treffer, allerdings können nicht so viele Zuordnungen vorgenommen werden.

Bei Eingabe von fünf Beispieldokumenten erhält man für das Touchpad-Profil sogar noch ein besseres Ergebnis: Mit einem Schwellwert von 24 Prozent können insgesamt 284 Patente neu klassifiziert werden. Den Wert höher zu setzen, hätte auch hier zur Folge, dass weniger Neuklassifizierungen erfolgen könnten. Und ein niedrigerer Wert würde wiederum zu viele Fehltreffer mit sich ziehen. Für das Fußgängerschutz-Profil sollte ein Wert von 30 Prozent gewählt werden. Damit werden 51 Patentdokumente klassifiziert. Bereits der Schwellwert darunter bringt schon mehr als die Hälfte an Fehlzuordnungen mit sich.

Und auch hier kann wieder eine wichtige Erkenntnis gemacht werden: Je besser die Beispieldokumente sind, umso mehr Patentdokumente können klassifiziert werden. Beim Touchpad-Profil werden mit einem Wert von 24 Prozent sehr viele Neuklassifikationen ermöglicht. Beim Fußgängerschutz-Profil ist schon mit einem Wert von 30 Prozent ein Optimum erreicht, es können lange nicht so viele Patentdokumente klassifiziert werden, da die Beispieldokumente bei Weitem nicht so gut sind, wie beim Touchpad-Profil. Je eindeutiger die Beispiele sind, desto besser wird also das Ergebnis.

Und schließlich macht sich bei der letzten Eingabemöglichkeit mit Hilfe des höchsten Precision-Werts auch eine wichtige Erkenntnis breit. Für das Touchpad-Profil wird hier das schlechteste Ergebnis von allen Untersuchungen aufgewiesen. Mit einem Schwellwert von 43 Prozent können lediglich 16 neue Patente dem T-Schlüssel zugeordnet werden. Bei einem niedrigeren Schwellwert zeigt sich deutlich, dass das Doppelte an Fehlklassifikationen zu Tragen kommt. Jedoch sind lediglich 16 mögliche Patentzuordnungen bei Weitem nicht zufrieden stellend. Bei genauerer Untersuchung der Eingabe stellt sich heraus, dass mit diesem Ergebnis hätte gerechnet werden müssen, da die Suchbegriffe „input instruction“ und „touch“ lauten. Diese beiden Begriffe haben zwar bei der Known-Item-Analyse den höchsten Precision-Wert erhalten, und es wurden 42 Known-Items von insgesamt 47 Gesamttreffern zurückgeliefert, bei der Schwellwertbestimmung jedoch zeigt sich, dass die beiden Begriffe schlecht sind. Im Ergebnisdiagramm der Anfrage wird folgendes Phänomen festgestellt: Zuerst werden richtige Treffer erhalten, insgesamt 16 Stück. Im Abstract der Treffer ist jedes Mal der Begriff „touch“ dabei. Ab Treffer-ID 17 jedoch erscheinen abwechselnd relevante und nichtrelevante Treffer. Das liegt unter anderem daran, dass der Begriff „input instruction“ mitverwendet wird. Der Begriff ist zu allgemein gehalten, da er nicht zwingend etwas mit dem Touchpad-Thema zu tun hat. Ein „input instruction“ kann für viele Themengebiete verwendet werden. In Verbindung mit „touch“ sind die beiden Begriffe jedoch eindeutig. Das bedeutet letzten Endes, dass sie in der Suchanfrage allerdings nicht gleichermaßen behandelt werden. Die Dokumente, in denen „touch“ vorkommt, werden als Treffer gewertet, und es gibt einen Ausschlag für einen relevanten Treffer, und die Dokumente, in denen „input instruction“ vorkommt, werden als Fehltreffer gewertet. Damit kommt der beobachtete Wechsel zwischen relevant und nichtrelevant zustande.

Das genaue Gegenteil zeigt die Untersuchung mittels des höchsten Precision-Werts für das Fußgängerschutz-Profil. Hier können mit einem Schwellwert von 17 Prozent insgesamt 150 Klassifikationen erfolgen. Das ist für dieses Profil die höchste Menge. Und auch der Schwellwert ist der niedrigste, der in der ganzen Evaluierung festgestellt werden kann. Ferner zeigt sich bei genauerer Untersuchung des Schlagwortes, dass dieses Ergebnis auch hätte erwartet werden müssen. Eingegeben wird „pedestrian detection“. Für das Fußgängerschutz-Profil ist dieser Begriff mehr als eindeutig. Der Begriff erscheint bei jedem neu klassifizierten Dokument. Das liegt daran, dass das der einzige zutreffende Begriff für Fußgängerschutz ist. Genauer und eindeutiger kann das Schlagwort nicht sein. Jedes Dokument in dem „pedestrian“ vorkommt, wird dementsprechend als Treffer gezählt.

Damit wird aber auch sehr deutlich, dass der Schwellwert nie mittels eines Schlagwortes bestimmt werden sollte, da die Ergebnisse so unterschiedlich ausfallen können. Wird ein eindeutiges Wort eingegeben, erfolgt ein Höchstmaß an Klassifikationen, ist der Begriff jedoch zu allgemein gehalten, werden nur sehr wenige Klassifikationen vorgenommen.

Empfohlen wird deshalb, Patentklassifikationen nur mittels Eingabe des T-Schlüssels oder mittels Eingabe von Beispieldokumenten vorzunehmen.

Weiterhin empfohlen wird daher, den Schwellwert für die verschiedenen Technologieschlüssel einzeln nach Profil festzulegen, damit ein Optimum an Patentdokumenten zugeordnet werden kann. Somit muss der Schwellwert immer wieder umgestellt werden. Das wird aber nötig, damit das Optimum ausgeschöpft werden kann. Das heißt aber letztendlich, dass die Schwellwertbestimmung themenabhängig ist. Einen allgemeingültigen Wert, der für alle T-Schlüssel übernommen werden könnte, gibt es nicht, und der Vorgang der Schwellwertbestimmung kann folglich nicht automatisiert werden. Der Wert muss jedes Mal von neuem „per Hand“ herausgefunden und eingestellt werden.

Für das Profil Laserschweißen hingegen lässt sich kein Schwellwert bestimmen. Die gewonnenen Ergebnisse in Form der Diagramme zeigen deutlich auf, dass es keine größeren Blöcke mit eindeutigen Treffern, geschweige denn Blöcke mit falschen Treffern gibt. Ein relevanter Treffer wechselt sich regelmäßig mit einem falschen ab. Hier tritt das Phänomen auf, dass bestimmte *matching words* zu oft vorkommen und folglich von der Relevanz her weniger gewichtet werden. Andere Begriffe sind dann wichtiger. Die Begriffe fungieren damit als Stoppwörter, und das WR-System interpretiert die dahinter stehenden Patentdokumente einmal als Treffer und einmal als Fehltreffer.

Durch die letzte Untersuchungsform soll schließlich bewertet werden, ob ein Einfluss der Internationalen Patentklassifikation erfolgt oder ob sie keinerlei Bedeutung für die Arbeitsweise des Text-Mining-Verfahrens hat. Erstaunlicherweise hat sie keinen großen Einfluss auf die Ergebnisse. Zwischen den jeweils dazu erstellten Diagrammen werden bei keinem der drei T-Schlüsselprofile große Abweichungen festgestellt. Offensichtlich ist jedoch, dass die Trefferzahl bei der Recherchemöglichkeit ohne Berücksichtigung der IPC um einiges größer ist und die ersten Fehltreffer, von der Relevanz her bewertet, weiter hinten stehen. Das lässt sich damit erklären, dass bei der ersten Testversion mit Berücksichtigung der IPC-Merkmale die IPC-Klasse sowie die reinen Schlagwörter als Stoppwörter fungieren. Bei der zweiten Möglichkeit jedoch sind nur die reinen Schlagwörter Stoppwörter - und letztendlich ergibt das mehr Treffer.
Des Weiteren tauchen die ersten Fehltreffer bei beiden Varianten an ähnlichen Stellen auf und sind nicht - wie anfangs vermutet - an gegensätzlichen Stellen aufzufinden.

Deshalb kann von einem sehr ähnlichen Endergebnis ausgegangen werden, und es ist somit im Endeffekt nicht relevant, ob die IPC berücksichtigt wird oder nicht.

7.2 Empfehlung

Die im Rahmen dieser Arbeit durchgeführte Evaluierung bringt einige bedeutende Ergebnisse.

Mittels des WR-Systems erfolgt eindeutig eine Erleichterung des Klassifikationsprozesses durch die Durchführung einer halbautomatischen Klassifikation. Wie gefordert werden die Zuordnungsstufen eins und zwei der derzeitigen Realisierung ersetzt, und auch eine rein manuelle Zuordnung ist durch die Automatisierung nicht mehr relevant. Somit werden Kosten und Zeit gespart. Des Weiteren wird erkannt, dass das WR-System nicht besser als die derzeitige Klassifikationsmöglichkeit ist, aber die gleiche Qualität aufweisen kann. Die Lösung der nicht zuordnungsfähigen 10 bis 20 Prozent an Patentdokumenten als Grund für die Suche nach einer Alternative hat sich somit allerdings nicht ergeben. Das heißt, dass sich die Fehlerraten der ePortfolio-Statistiken nicht ändern werden.

Das wichtigste und ausschlaggebendste Ergebnis ist jedoch, dass das WR-System eine einfachere Klassifikation ermöglicht. Und darin liegt der entscheidende Punkt.

So kann alleine mit fünf Beispieldokumenten eine größere Anzahl an relevanten Patentdokumenten gefunden werden, als das mit einem anderen Recherchesystem überhaupt möglich ist. Der derzeitige Aufwand verringert sich erheblich, und deshalb verhilft What's Related als Komplettsystem - und nicht von den einzelnen Funktionen her betrachtet - zu deutlicher Erleichterung. Ferner erfolgt auch eine schnellere Informationsaufnahme, da auch unerfahrene Benutzer leicht mit dem System arbeiten können und schnell einen inhaltlichen Überblick über ein bestimmtes Themengebiet erhalten.

Auf Grund dessen wird eine klare Empfehlung für das System ausgesprochen.

7.3 Ausblick

Die weiteren Entwicklungsmöglichkeiten werden in einer umfangreichen Benutzerverwaltung gesehen. Eine Idee für eine mögliche Oberflächenlösung besteht bereits. Diese Alternative soll allerdings nur für die Mitarbeiter gelten, die die Klassifikationen der T-Schlüssel vornehmen. Für die Recherche an sich und für die Benutzer des Systems müsste eine neue Weboberfläche realisiert werden.

Bei der besagten Lösung handelt es sich um eine dateibasierte Methode. Ein Dateiverzeichnis entspricht dabei einem T-Schlüssel. Die Ergebnisse werden schließlich alle als Verzeichnisstruktur aufgelistet. Möglich ist dadurch ein schnelleres Durchsehen und Abarbeiten der Patentdokumente. Die Verzeichnisstruktur dient dabei als Lernmenge und wird von der Stelle aus bearbeitet. Die anschließend bearbeiteten und gesäuberten Daten werden alsdann in eine neue Dateistruktur übernommen. Vorteile, die sich daraus ergeben, sind zum einen eine Erleichterung der Arbeit und zum anderen die Möglichkeit eines effizienteren Arbeitens im Vergleich zu einem Webinterface, wie es in der Diplomarbeit verwendet wurde. Des Weiteren kann der Benutzer mit seiner

Bearbeitung zu jeder Zeit aufhören und zu einem Zeitpunkt seiner Wahl genau an derselben Stelle weitermachen, an der er aufgehört hat. Der Nachteil, der sich daraus ergibt besteht darin, dass es etwas länger dauert, die Dateien zu erzeugen.

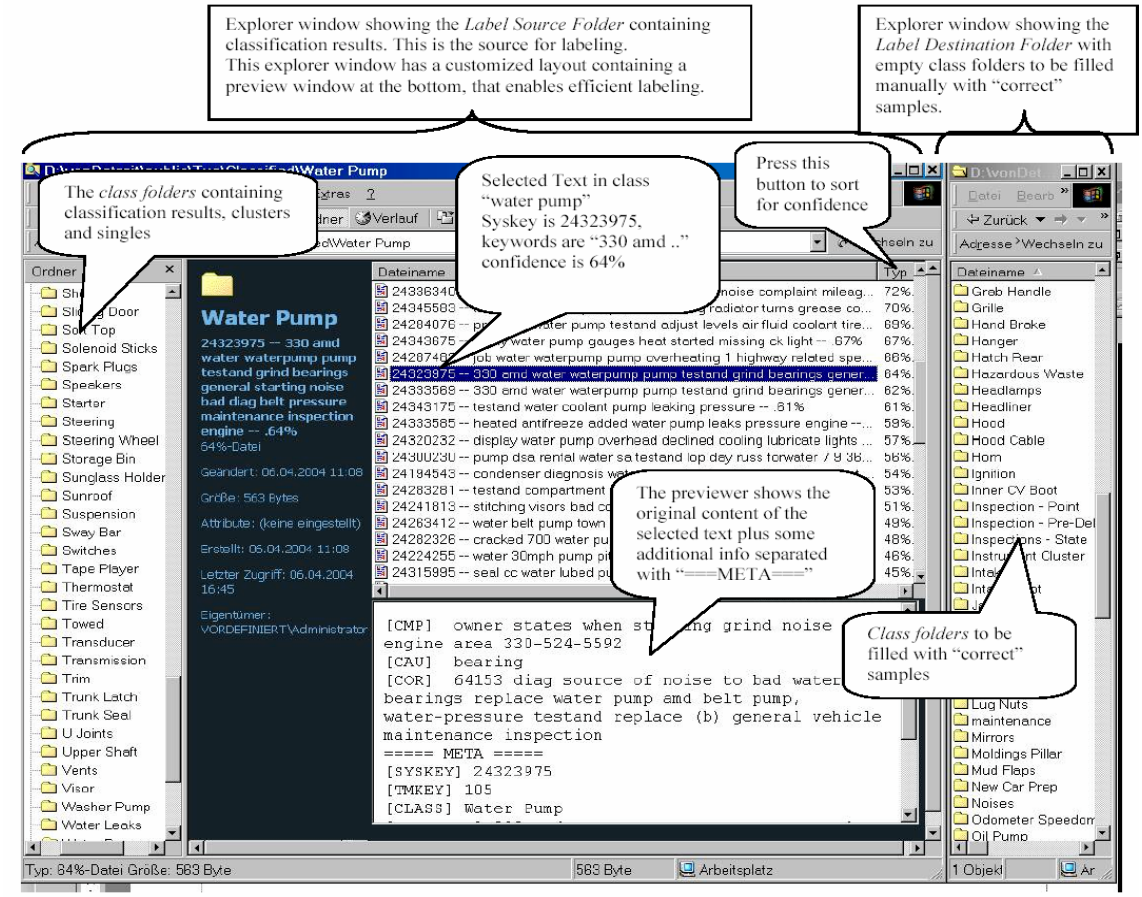


Abbildung 44: Mögliches Interface[117]

Abbildung 44 stellt ein mögliches Interface dar. In der linken Hälfte steht die zu bearbeitende Lernmenge in den einzelnen Ordnern. In der Mitte werden die neu erstellten Klassifikationen für einen T-Schlüssel aufgelistet. Im unteren Teil besteht die Möglichkeit einer Vorschaufunktion, und ganz rechts stehen die fertigen Ordner, die mit den bearbeiteten Dokumenten aus der Mitte befüllt werden.
Im weiteren Verlauf soll das WR-System mit dem kompletten IPM/C-Datenbestand befüllt werden. Das Ziel für die Zukunft besteht darin, dass die zu integrierenden wöchentlichen Datenbank-Updates mit Hilfe des Text-Mining-Verfahrens die entsprechenden Technologieschlüssel zugeordnet bekommen und anschließend in die Projektdatenbank geladen werden. Somit wird der Durchlauf durch die BRS-Datenbank umgangen.

Schließlich muss als letzte Hürde die Integration des Systems in IPM/C durchgeführt werden.

[117] Quelle: Bohnacker/Morandell (2004), S. 3

8 Zusammenfassung

Innovative Informationstechnologien wie das Text-Mining ersetzen nicht den Menschen, sondern unterstützen ihn vielmehr. Das hat sich auch in der vorliegenden Arbeit gezeigt.

Ausgehend von einer bestehenden Problemsituation innerhalb der Patentabteilung der DaimlerChrysler AG sind Anforderungen an eine mögliche Lösungsalternative in Form eines statistischen Text-Mining-Verfahrens gestellt worden. Nach einer Überprüfung der Möglichkeiten, die das Verfahren bietet, sind umfangreiche Tests zur Untersuchung der Aufgabenerfüllung durchgeführt worden. Die Evaluierung ist dabei nach festgelegten Kriterien vorgenommen worden. Dazu wurden zum einen die grundlegenden Funktionalitäten getestet, und zum anderen wurden in einem weiteren Schritt spätere Konfigurationsmöglichkeiten festgelegt.

Als Ergebnis der Untersuchung konnte eine Empfehlung für einen zukünftigen Einsatz des Verfahrens ausgesprochen werden.

Anhang A: Ergebnisse der Known-Item-Analyse

Im Folgenden sind die einzelnen Ergebnistabellen der Recherche aus Kapitel 6.2.2.2 für das Touchpad-Profil aufgeführt. Zur besseren Darstellung werden die Eingabewerte in Tabelle 9 einer fortlaufenden Nummer zugeordnet. In den Ergebnistabellen sind der Eingabewert, das Ergebnis, bestehend aus Gesamttrefferzahl und der getroffenen Known-Items, und die Auswertung in Prozent, bestehend aus berechneter Precision und Availability, ersichtlich.

Nummerierung	Art	Eingabewert
1	T-Schlüssel	6.4.10.3.1.0.V14
2	PAN	2003-727746
3		1999-371928
4		2002-324577
5		1998-500686
6		2003-117243
7		2000-045290
8		2000-141707
9		2000-495815
10	Begriffe	touch pad
11		touch panel
12		touch screen
13		touch
14		user`s finger touch
15		touch surface
16		touch sensor
17		input instruction

Tabelle 9: Zuordnungen[118]

[118] Quelle: Eigene Darstellung

A.1 Recherche nach Patenten ohne T-Schlüssel

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1	4851	12	0,25	100,00
2	1277	11	0,86	91,67
3	1524	2	0,13	16,67
4	1432	12	0,84	100,00
5	1779	10	0,56	83,33
6	2240	9	0,40	75,00
10	1939	11	0,57	91,67
11	2000	12	0,60	100,00
12	2001	12	0,60	100,00
13	1324	12	0,91	100,00
14	1569	11	0,70	91,67
15	1454	12	0,83	100,00
16	1981	12	0,61	100,00

Tabelle 10: Ergebnisse der einfachen Suche[119]

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1, 2	1229	12	0,98	100,00
1, 3	1385	12	0,87	100,00
1, 4	1241	12	0,97	100,00
1, 5	1178	12	1,02	100,00
1, 6	1998	12	0,60	100,00
1, 6, 3	1813	12	0,66	100,00
1, 6, 3, 4	1709	12	0,70	100,00
1, 6, 3, 4, 5	1710	12	0,70	100,00
1, 13	1752	12	0,68	100,00
1, 11	1719	12	0,70	100,00
1, 12	1575	12	0,76	100,00
1, 10	1348	12	0,89	100,00
2, 3	1837	10	0,54	83,33
2, 3, 4	1969	11	0,56	91,67
2, 3, 4, 5	1765	12	0,68	100,00

[119] Quelle: Eigene Darstellung

2, 3, 4, 5, 6	1769	12	0,68	100,00
2, 10	822	10	1,22	83,33
2, 3, 10	872	10	1,15	83,33
2, 3, 4, 10	885	11	1,24	91,67
2, 3, 4, 5, 10	925	11	1,19	91,67
2, 3, 4, 5, 6, 10	900	11	1,22	91,67
2, 13	1386	12	0,87	100,00
2, 3, 13	1524	12	0,79	100,00
2, 3, 4, 13	1448	12	0,83	100,00
2, 3, 4, 5, 13	1368	12	0,88	100,00
2, 3, 4, 5, 6, 13	1370	12	0,88	100,00
17, 13	1811	11	0,61	91,67

Tabelle 11: Ergebnisse der kombinierten Suche für den 1.Fall[120]

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1, 7	1250	12	0,96	100,00
1, 8	1512	12	0,79	100,00
1, 9	1132	12	1,06	100,00
1, 2, 4, 7	1368	12	0,88	100,00
7	1272	12	0,94	100,00
8	1804	10	0,55	83,33
9	1417	12	0,85	100,00
7, 13	1355	12	0,89	100,00
8, 13	1623	12	0,74	100,00
7, 8, 9	1592	12	0,75	100,00

Tabelle 12: Ergebnisse der kombinierten Suche für den 2.Fall[121]

[120] Quelle: Eigene Darstellung
[121] Quelle: Eigene Darstellung

A.2 Recherche nach Patenten mit T-Schlüssel

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1	149	45	30,20	100,00
2	57	38	66,67	84,44
3	49	26	53,06	57,78
4	95	37	38,95	82,22
5	139	36	25,90	80,00
6	98	39	39,80	86,67
10	92	40	43,48	88,89
11	215	41	19,07	91,11
12	117	43	36,75	95,56
13	52	40	76,92	88,89
14	57	41	71,93	91,11
15	53	40	75,47	88,89
16	251	40	15,94	88,89

Tabelle 13: Ergebnisse der einfachen Suche[122]

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1, 2	58	42	72,41	93,33
1, 3	58	42	72,41	93,33
1, 4	74	42	56,76	93,33
1, 5	64	42	65,63	93,33
1, 6	76	42	55,26	93,33
1, 6, 3	69	42	60,87	93,33
1, 6, 3, 4	80	43	53,75	95,56
1, 6, 3, 4, 5	90	43	47,78	95,56
1, 13	65	44	67,69	97,78
1, 11	96	44	45,83	97,78
1, 12	66	44	66,67	97,78
1, 10	60	44	73,33	97,78
2, 3	73	40	54,79	88,89
2,3, 4	75	41	54,67	91,11

[122] Quelle: Eigene Darstellung

2, 3, 4, 5	67	43	64,18	95,56
2, 3, 4, 5, 6	100	43	43,00	95,56
2, 10	40	35	87,50	77,78
2, 3, 10	48	41	85,42	91,11
2, 3, 4, 10	49	41	83,67	91,11
2, 3, 4, 5, 10	50	42	84,00	93,33
2, 3, 4, 5, 6, 10	50	42	84,00	93,33
2, 13	53	42	79,25	93,33
2, 3, 13	54	43	79,63	95,56
2, 3, 4, 13	54	43	79,63	95,56
2, 3, 4, 5, 13	54	43	79,63	95,56
2, 3, 4, 5, 6, 13	54	43	79,63	95,56
17, 13	47	42	89,36	93,33

Tabelle 14: Ergebnisse der kombinierten Suche für den 1. Fall[123]

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1, 7	56	42	75,00	93,33
1, 8	74	42	56,76	93,33
1, 9	55	42	76,36	93,33
1, 7, 13	56	43	76,79	95,56
1, 8, 13	66	43	65,15	95,56
1, 9, 13	54	43	79,63	95,56
1, 2, 4, 7	63	42	66,67	93,33
7	47	35	74,47	77,78
8	87	22	25,29	48,89
9	43	36	83,72	80,00
7, 13	51	40	78,43	88,89
8, 13	55	40	72,73	88,89
7, 8, 9	63	38	60,32	84,44

Tabelle 15: Ergebnisse der kombinierten Suche für den 2.Fall[124]

[123] Quelle: Eigene Darstellung
[124] Quelle: Eigene Darstellung

A.3 Recherche nach Patenten mit oder ohne T-Schlüssel

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1	5000	56	1,12	98,25
2	1293	43	3,33	84,21
3	1517	44	2,90	77,19
4	1413	48	3,40	78,95
5	1807	44	2,43	0,00
6	2473	45	1,82	89,47
10	1967	51	2,59	89,47
11	2001	51	2,55	89,47
12	2001	51	2,55	89,47
13	1376	51	3,71	89,47
14	1568	43	2,74	75,44
15	1507	51	3,38	89,47
16	1983	52	2,62	91,23

Tabelle 16: Ergebnisse der einfachen Suche[125]

Eingabe	Ergebnis		Auswertung in %	
	Treffer	Gesamtmenge	Precision	Availability
1, 2	1279	54	4,22	94,74
1, 3	1434	54	3,77	94,74
1, 4	1289	54	4,19	94,74
1, 5	1216	54	4,44	94,74
1, 6	2046	54	2,64	94,74
1, 6, 3	1854	54	2,91	94,74
1, 6, 3, 4	1736	54	3,11	94,74
1, 6, 3, 4, 5	1744	55	3,15	96,49
1, 13	1817	54	2,97	94,74
1, 11	1769	54	3,05	94,74
1, 12	1621	55	3,39	96,49
1, 10	1401	54	3,85	94,74
2, 3	1828	48	2,63	84,21
2,3, 4	1689	50	2,96	87,72

[125] Quelle: Eigene Darstellung

2, 3, 4, 5	1764	52	2,95	91,23
2, 3, 4, 5, 6	1757	54	3,07	94,74
2, 10	853	41	4,81	71,93
2, 3, 10	912	47	5,15	82,46
2, 3, 4, 10	919	46	5,01	80,70
2, 3, 4, 5, 10	956	46	4,81	80,70
2, 3, 4, 5, 6, 10	936	47	5,02	82,46
2, 13	1439	54	3,75	94,74
2, 3, 13	1578	55	3,49	96,49
2, 3, 4, 13	1502	55	3,66	96,49
2, 3, 4, 5, 13	1422	55	3,87	96,49
2, 3, 4, 5, 6, 13	1424	55	3,86	96,49
17, 13	1831	44	2,40	77,19

Tabelle 17: Ergebnisse der kombinierten Suche für den 1. Fall[126]

Eingabe	Ergebnis		Auswertung in %	
	Gesamttreffer	Known-Items	Precision	Availability
1, 7	1304	54	4,14	100,00
1, 8	1537	54	3,51	100,00
1, 9	1187	54	4,55	100,00
1, 7, 13	1468	55	3,75	101,85
1, 8, 13	1799	55	3,06	101,85
1, 9, 13	1377	55	3,99	101,85
1, 2, 4, 7	1407	54	3,84	94,74
7	1298	46	3,54	80,70
8	1825	22	1,21	38,60
9	1448	48	3,31	84,21
7, 13	1406	52	3,70	91,23
8, 13	1678	52	3,10	91,23
7, 8, 9	1622	47	2,90	82,46

Tabelle 18: Ergebnisse der kombinierten Suche für den 2. Fall[127]

[126] Quelle: Eigene Darstellung
[127] Quelle: Eigene Darstellung

Anhang B: Ergebnisse der Schwellwertbestimmung

Im Folgenden werden die verschieden Schwellwertbestimmungen für das Fußgängerschutz-Profil vorgenommen.

B.1 Suchanfrage bei Eingabe des Technologieschlüssels

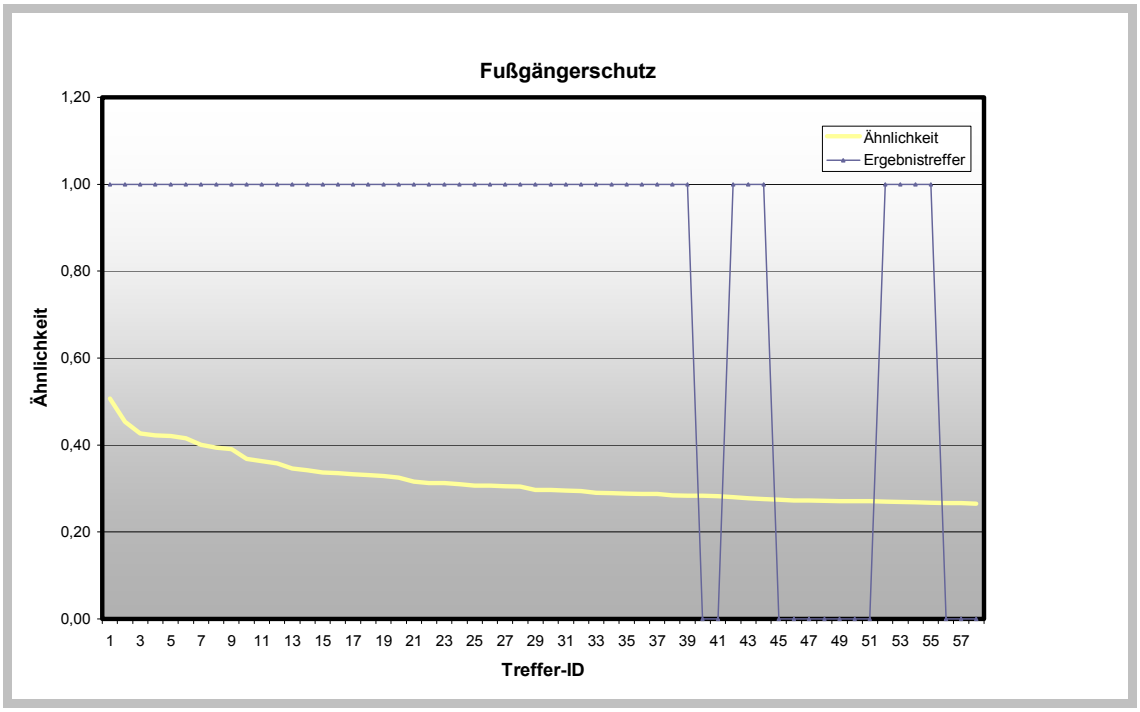


Abbildung 45: Recherche mit T-Schlüssel in der Testkollektion bei $W_3 = 27\%$ [128]

❖ Was ist zu sehen?

Erkennbar ist, dass in Diagramm 45 innerhalb der Testmenge mit einem eingestellten Schwellwert von 27 Prozent zuerst ein großer Block mir richtig klassifizierten Patentdokumenten gefunden wird. Die ersten Fehltreffer erscheinen bei einem Ähnlichkeitswert von 28 Prozent bei Treffer-ID 40 und 41. Insgesamt werden 59 Dokumente getroffen. Davon sind 46 eindeutig richtig und 13 falsch zugeordnet. Nach den ersten beiden Fehltreffern tauchen abermals drei relevante Treffer auf. Die darauf folgenden Treffer sind jedoch schon nicht mehr richtig klassifiziert.

[128] Quelle: Eigene Darstellung

❖ Was wird daraus geschlossen?

Für das gesuchte Ergebnis bedeutet das aber letztendlich, dass ein Schwellwert von 29 Prozent ein Optimum an richtig klassifizierten und relevanten Patenten herausbringt. Damit können 37 mögliche Klassifizierungen vorgenommen werden. Die zwei Treffer Differenz bleiben unberücksichtigt, da sie bereits zu der „abgeschnittenen" Menge gehören. Von den ursprünglich 53 bekannten Treffern aus der Testkollektion können somit mehr als die Hälfte klassifiziert werden.

B.2 Suchanfrage bei Eingabe mehrerer Beispieldokumente

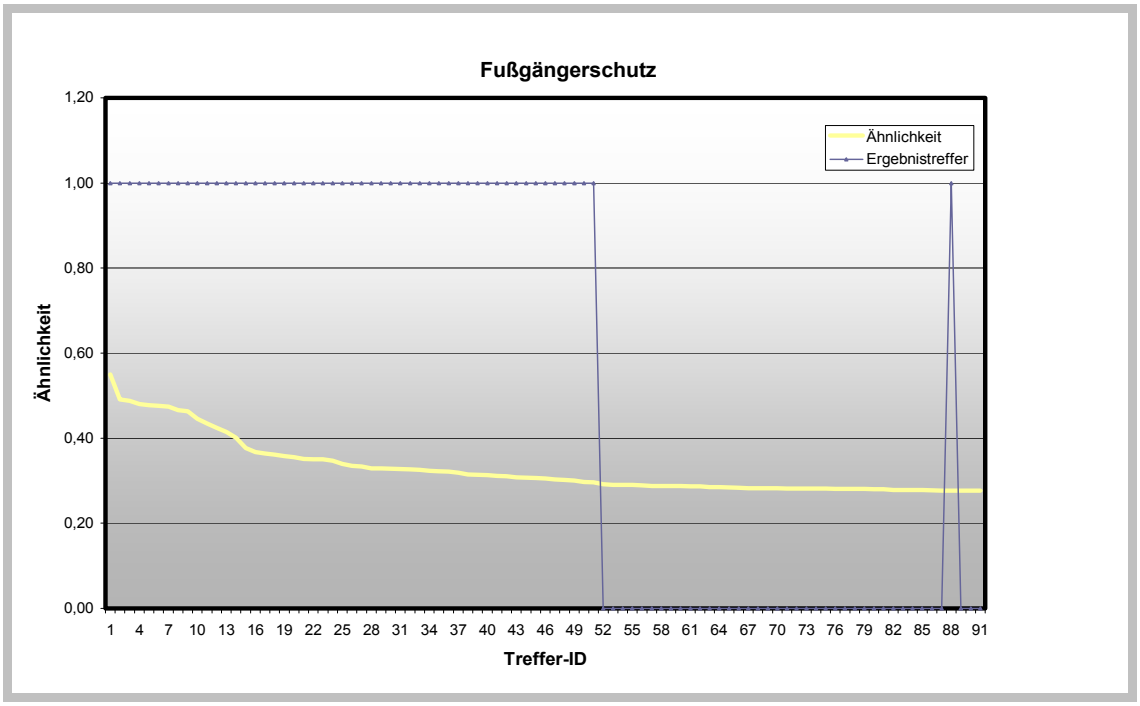


Abbildung 46: Recherche mit Dokumenten in der Testkollektion bei $W_1 = 28\%$ [129]

❖ Was ist zu sehen?

Erkennbar ist bei der Untersuchung der Testkollektion mit einem Schwellwert von 28 Prozent, dass zuerst ein großer Block mit richtig klassifizierten Patentdokumenten gefunden wird. Ab dem ersten Fehltreffer mit der Nummer 52 werden keine richtigen Klassifikationen mehr vorgenommen. Der Ähnlichkeitswert liegt hier bei 29 Prozent. Alle Treffer, die noch folgen, sind Fehltreffer. Insgesamt sind 92 Gesamttreffer ersichtlich, mit 32 Fehlzuordnungen und 61 relevanten Treffern.

❖ Was wird daraus geschlossen?

Aus dem Diagramm lässt sich damit ein Schwellwert von 30 Prozent festlegen. Mit diesem Endwert können 51 mögliche Klassifizierungen vorgenommen werden.

[129] Quelle: Eigene Darstellung

B.3 Suchanfrage bei Eingabe des höchsten Precision-Werts

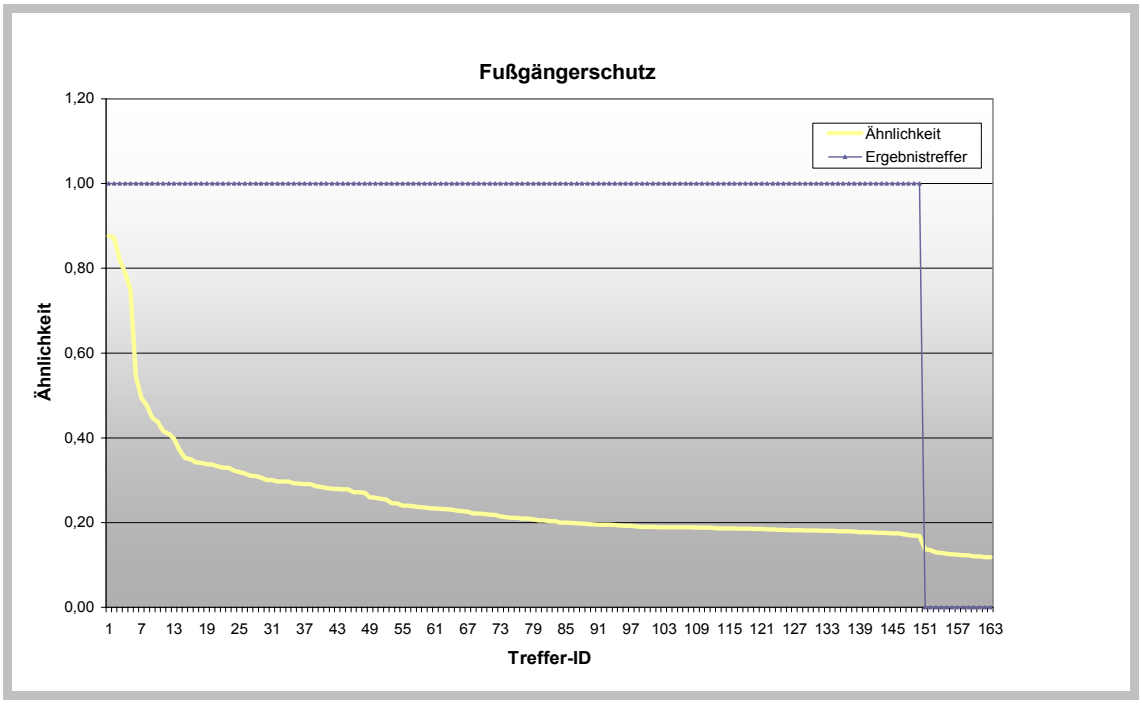


Abbildung 47: Recherche mit Begriffen in der Testkollektion bei $W_4 = 12\%$ [130]

❖ Was ist zu sehen?

In Abbildung 47 wird ein Schwellwert von 12 Prozent überprüft. Und auch wie in den vorigen beiden Diagrammen ist zuerst ein großer Block mit richtig klassifizierten Patentdokumenten für die Testmenge zu sehen. Der erste Fehltreffer erscheint bei einem Ähnlichkeitswert von 14 Prozent bei Treffer-ID 151. Danach tauchen lediglich falsch zugeordnete Patentdokumente bei einer Gesamttrefferzahl von 164 Dokumenten auf. Alles in allem werden 150 richtige und 14 Fehlklassifikationen festgestellt.

❖ Was wird daraus geschlossen?

Für das gesuchte Ergebnis bedeutet das, dass ein Schwellwert von 17 Prozent ein Optimum an richtig klassifizierten und relevanten Patenten herausbringt. Damit können 150 Dokumente dem Fußgängerschutz-Profil zugeordnet werden, was um einiges mehr ist, als die bekannten 53 Patente.

B.4 Zusammenfassung

In der nachfolgenden Tabelle werden die Ergebnisse der Schwellwertbestimmung für das Fußgängerschutz-Profil zusammengefasst.

[130] Quelle: Eigene Darstellung

Lernkollektion				
Schwellwert	**Known Items**	**Richtige Treffer**	**Falsche Treffer**	**Prozent**
Eingabe des T-Schlüssels				
10%	309	307	162	99,4
1. Fehltreffer: 37%	309	202	1	65,4
2. Fehltreffer: 29%	309	262	2	84,8
3. Fehltreffer: 27%	309	271	3	87,7
Eingabe von fünf Beispieldokumenten				
10%	309	308	302	99,7
1. Fehltreffer: 38%	309	185	1	59,9
Eingabe des höchsten Precision-Werts				
10%	309	177	68	57,3
1. Fehltreffer: 19%	309	128	1	41,4
2. Fehltreffer: 17%	309	147	2	47,6
3. Fehltreffer: 12%	309	149	3	48,2
Testkollektion				
Eingabe des T-Schlüssels				
10%	Menge unbekannt, mindestens 53	72	4.459	135,8
37%	Menge unbekannt, mindestens 53	10	0	18,9
29%	Menge unbekannt, mindestens 53	37	0	69,8
27&	Menge unbekannt, mindestens 53	46	13	86,8
Ergebnis: Schwellwert von 29% ergibt 37 neue Klassifizierungen				
Eingabe von fünf Beispieldokumenten				
10%	Menge unbekannt, mindestens 53	71	2.440	134
28%	Menge unbekannt, mindestens 53	61	32	115,1
Ergebnis: Schwellwert von 30% ergibt 51 neue Klassifizierungen				
Eingabe des höchsten Precision-Werts				
10%	Menge unbekannt, mindestens 53	50	447	94,3
19%	Menge unbekannt, mindestens 53	118	0	222,6
17%	Menge unbekannt, mindestens 53	150	0	283
12%	Menge unbekannt, mindestens 53	150	14	283
Ergebnis: Schwellwert von 17% ergibt 150 neue Klassifizierungen				

Tabelle 19: Zusammenfassung der Ergebnisse für das Fußgängerschutz-Profil[131]

[131] Quelle: Eigene Darstellung

Zuerst werden die Untersuchungen für die Lernkollektion bei Eingabe des T-Schlüssels, der fünf Beispieldokumente und bei Eingabe der Begriffe mit der höchsten Precision dargestellt. Des Weiteren wird in dem Zusammenhang angegeben, wie viele Known-Items innerhalb der beiden Kollektionen vorkommen und wie viele davon bei einem bestimmten Schwellwert getroffen werden. Das entspricht dem Feld „Richtige Treffer“. Ein falscher Treffer wird vermerkt, wenn innerhalb der Treffermenge keine Known-Items vorkommen. Das letzte Feld entspricht dem Prozentsatz, mit dem die richtigen Known-Items erkannt werden. Als Erstes werden die Ergebnisse für den momentanen Schwellwert von 10 Prozent aufgeführt. Anschließend erfolgt die Überprüfung der ersten Fehltreffer. Die Schwellwerte der Fehltreffer werden im Weiteren für die Testkollektion überprüft. Für diese Kollektion ist dabei nur die Mindestmenge der Known-Items bekannt, nicht aber, wie viele es insgesamt innerhalb der Testmenge gibt. Auf Grund dessen kann für den derzeitigen Schwellwert nur eine ungefähre Angabe gemacht werden. Aus den getroffenen Erkenntnissen der Testkollektion kann dann letztendlich der Schwellwert bestimmt werden.

Anhang C: Einfluss der IPC

Im Folgenden wird der Einfluss der Internationalen Patentklassifikation auf das System What's Related untersucht. Für die Untersuchung werden die beiden Profile Laserschweißen und Touchpad verwendet. Zur besseren Veranschaulichung werden die Diagramme auf der nächsten Seite dargestellt.

C.1 Das Laserschweißen-Profil

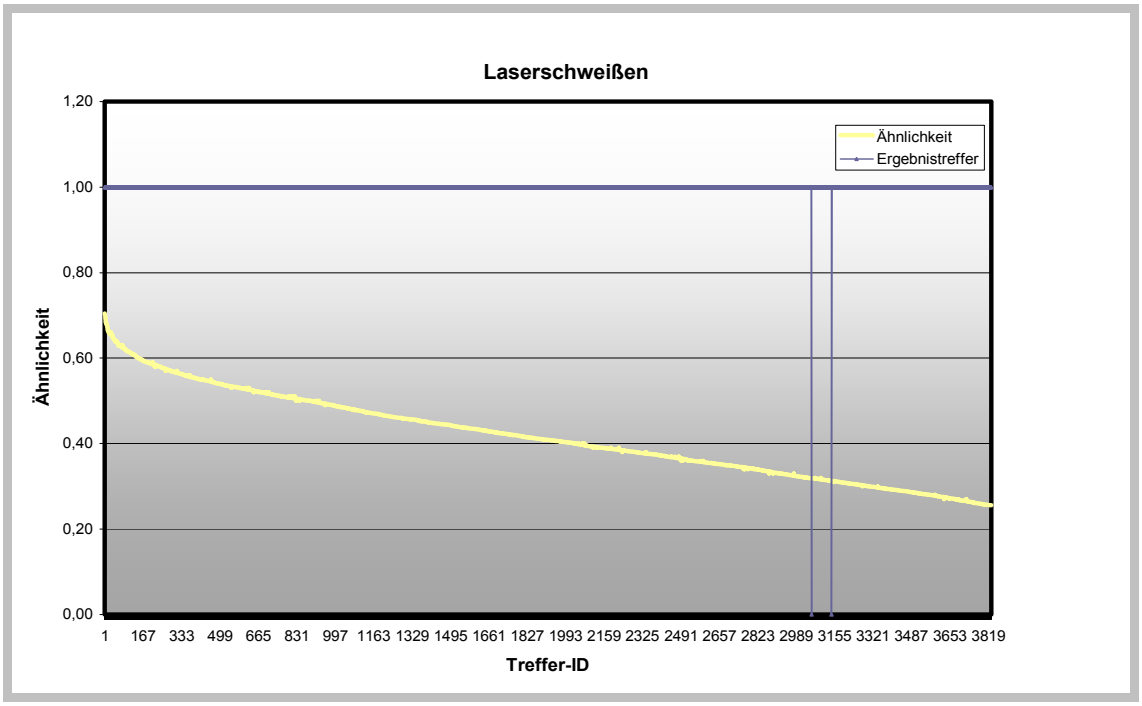


Abbildung 48: Recherche in der Lernkollektion bei $S_{Default} = 10\%$ mit IPC[132]

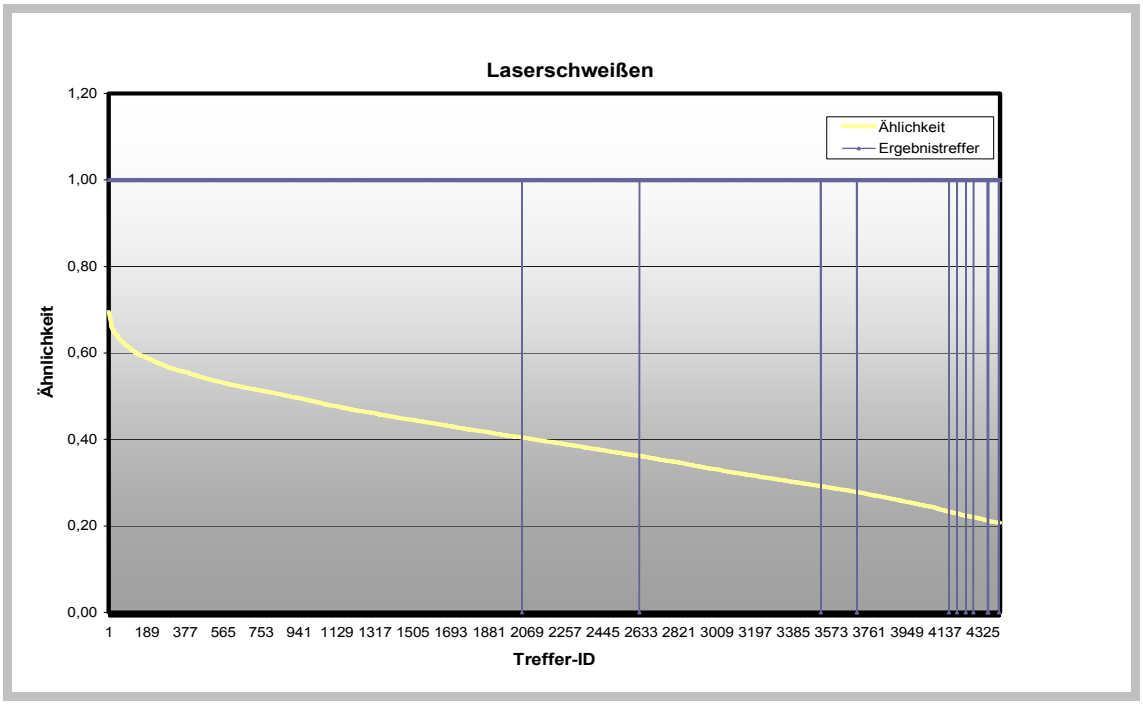


Abbildung 49: Recherche in der Lernkollektion bei $S_{Default} = 10\%$ ohne IPC[133]

[132] Quelle: Eigene Darstellung
[133] Quelle: Eigene Darstellung

❖ Was ist zu sehen?

In Abbildung 49 zeigt sich erneut eine höhere Treffergesamtzahl. Untersucht wird Anhang C mit einem Schwellwert von 10 Prozent, damit so viele Treffer wie möglich gemacht werden können. Die ersten Fehltreffer erscheinen in Diagramm 48 bei Treffer-ID 3.054 und 3.141. Davor sind nur richtig klassifizierte Dokumente anzutreffen. Nach den beiden einzigen Fehltreffern in einer Gesamtmenge von 3.830 Patentdokumenten erscheinen nur noch richtige Treffer. Insgesamt können dabei 3828 richtige Laserschweißen-Profile und 8 unpassende gefunden werden. In Abbildung 49 dagegen erscheint der erste Fehltreffer bei Treffer-ID 2.044. Insgesamt werden neun Fehltreffer bei 4.407 Gesamttreffern vorgefunden. Und auch die Gesamtzahl ist hier so aufgeteilt, dass 4395 richtige Dokumente und 12 falsche zugeordnet werden.

❖ Was wird daraus geschlossen?

Die beiden Diagramme weisen zwar eine Differenz bei den Fehltreffern auf, da sich die Trefferzahlen jedoch im Tausenderbereich bewegen, ist der Unterschied nicht allzu groß.

C.2 Das Touchpad-Profil

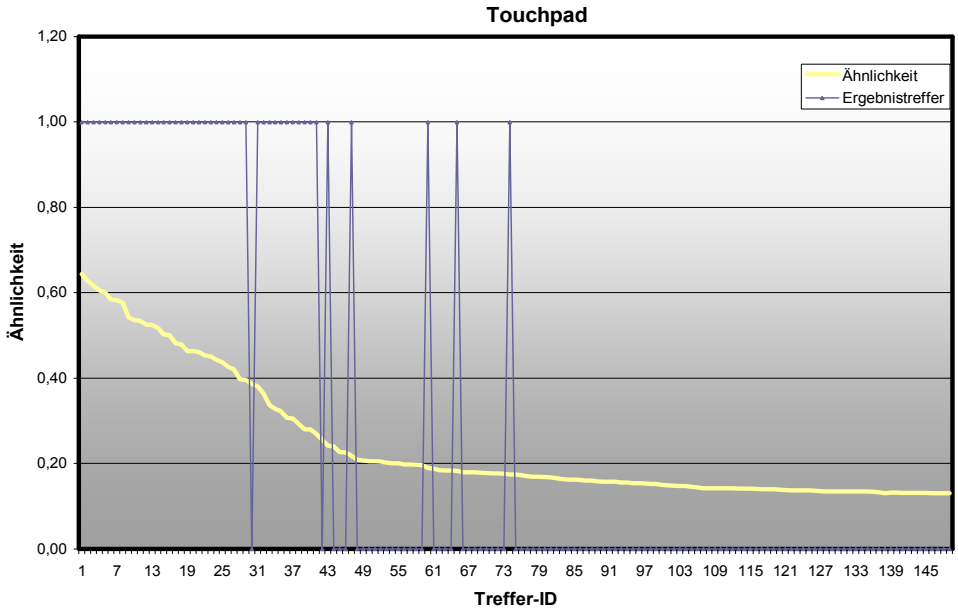


Abbildung 50: Recherche in der Lernkollektion bei $S_{Default} = 10\%$ mit IPC[134]

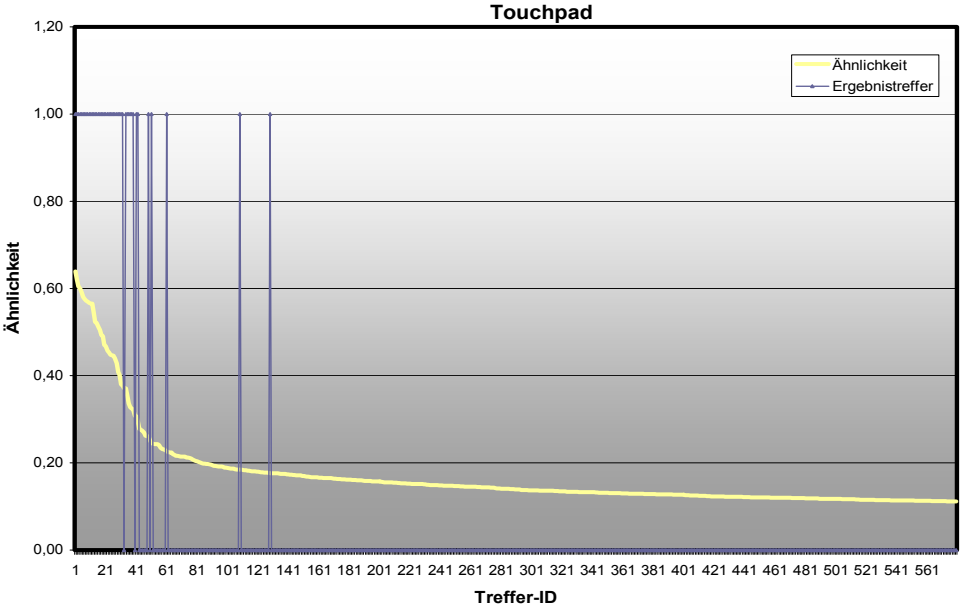


Abbildung 51: Recherche in der Lernkollektion bei $S_{Default} = 10\%$ ohne IPC[135]

[134] Quelle: Eigene Darstellung
[135] Quelle: Eigene Darstellung

❖ Was ist zu sehen?

In Abbildung 51 zeigt sich erneut eine höhere Treffergesamtzahl mit insgesamt 580 Patenten. Der erste Fehltreffer erscheint bei Treffer-ID 33. Davor sind nur richtig klassifizierte Dokumente anzutreffen. Insgesamt sind 45 Patente richtig und 535 falsch zugeordnet. In Abbildung 50 dagegen erscheint der erste Fehltreffer bei Nummer 30 bei einer Gesamtzahl von 149 Treffern, wovon 104 falsch zugeordnet sind. Somit besteht lediglich eine Differenz von drei Treffern bezüglich des ersten Fehltreffers.

❖ Was wird daraus geschlossen?

Das heißt, für die Recherche ohne Berücksichtigung der IPC können drei Dokumente mehr klassifiziert werden, und auch die Gesamttrefferzahl ist um einiges höher, als bei der Variante mit Berücksichtigung der Internationalen Patentklassifikation.

Literaturverzeichnis

Ackermann, Markus (2002): Lemmatisierung und Term-Clustering-Methoden zur Merkmalsgewinnung im Text-Mining. Diplomarbeit an der Universität Ulm, Fakultät für Informatik, Ulm, Deutschland.

Akademie.de asp GmbH (2004): net-lexikon. http://www.net-lexikon.de (Datum des Zugriffs: 05. Juli 2004).

Bartölke, Ingrun-Ulla (2000): Strategische Gruppen und Strategieforschung: Ansatz für eine dynamische Wettbewerbsbetrachtung. Betriebswirtschaftlicher Verlag Dr. Th. Gabler GmbH und Deutscher Universitäts-Verlag GmbH, Wiesbaden, Deutschland.

Bendl, Dr. Dr. Ernst; Weber, Georg (2002): Patentrecherche und Internet. Carl Heymanns Verlag KG, Köln, Deutschland.

Bohnacker, Ulrich; Dehning, Lars; Franke, Jürgen; Renz, Dr. Ingrid, Schneider, René (2000): Weaving Intranet Relations – Managing Web Content. E-Mail von Herrn Bohnacker.

Bohnacker, Ulrich; Morandell, Thomas (2004): Users Manual for Labeling and Maintaining the Classifier. E-Mail von Herrn Bohnacker.

Brockhaus (1997): Die Enzyklopädie. Band 6, 20. Auflage. Leipzig, Deutschland.

Däbritz, Dr. Erich (2001): Patente: Wie versteht man sie? Wie bekommt man sie? Wie geht man mit ihnen um? 2. Auflage. Verlag C. H. Beck oHG, München, Deutschland.

DaimlerChrysler AG, Communications (Hrsg.) (2002): Schürfen statt Suchen: Die Strategie für effizientes Suchen und Sortieren. Hightech Report: Berichte aus Forschung, Technik und Umwelt. Stuttgart, Band 1, S. 36-37.

DaimlerChrysler AG (2003): Intellectual Property Management, Electronic Profiles. http://intra-patent.daimlerchrysler.com/epps/. (Datum des Zugriffs: 25. März 2004).

Dörre, Jochen; Gerstl, Peter; Seiffert, Roland (2001): Volltextsuche und Text Mining. In: Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Rolf; Langer, Hagen (Hrsg.): Computerlinguistik und Sprachtechnologie: eine Einführung. Spektrum Akademischer Verlag, Heidelberg, Deutschland, S. 425-441.

Dresel, Robin; Hörnig, Daniel; Kaluza Harald; Peter, Anja; Roßmann, Nicola; Sieber, Wolfram (2001): Evaluation deutscher Web-Suchwerkzeuge: ein vergleichender Retrievaltest. Nfd Information - Wissenschaft und Praxis, Nr. 52. Darmstadt, Deutschland, S. 381-392.

Eisenrith, Eduard (1981): Das Patentwesen als Informationsquelle für Innovationen. Fortschrittsberichte der VDI Zeitschriften. Reihe: 10, Technik und Wirtschaft, Zeitschrift Nr. 10, VDI-Verlag GmbH, Düsseldorf, Deutschland.

Erdmann, Jérome (2002): Usability Engineering, Usabilty Evaluationsmethoden, Usability Probleme. Praxisarbeit. https://www.mnet.markant.de/support/studien-arbeit_usability.pdf. (Datum des Zugriffs: 25. März 2004).

Fendt, Dr. Heinrich (1991): Strategische Patentanalyse: Wettbewerbsvorteile durch Patentinformation. http://www.wi.fh-flensburg.de/wi/fendt/patent8.htm. (Datum des Zugriffs: 25. Mai 2004).

Ferber, Dr. Reginald (2003): Information Retrieval: Suchmodelle und Data-Mining-Verfahren für Textsammlungen und das Web. 1. Auflage. dpunkt.verlag GmbH, Heidelberg, Deutschland.

Ferber, Dr. Reginald (2003): Precision-Recall-Diagramm. http://information-retrieval.de/irb/Abbildung_40.html. (Datum des Zugriffs: 15. Juli 2004).

Fraunhofer-Patentstelle (2004): Die Bayerische Hochschul-Patentinitiative. http://www.pst.fhg.de/bayernpatent/glossar/index.html. (Datum des Zugriffs: 05. Juli 2004).

Gaus, Dr. Wilhelm (2003): Dokumentation und Ordnungslehre: Theorie und Praxis des Information Retrieval. 4. Auflage. Springer-Verlag, Berlin, Heidelberg, Deutschland.

Greif, Sigrid; Potkowik, Georg (1990): Patente und Wirtschaftszweige: Zusammenführung der Internationalen Patentklassifikation und der Systematik der Wirtschaftszweige. Carl Heymanns Verlag KG, Köln, Berlin, Bonn, München, Deutschland.

Gronau, Norbert (2001): Industrielle Standardsoftware – Auswahl und Einführung. Oldenbourg Verlag, München, Deutschland.

Haag, Markus (2002):Automatic Text Summarization: Evaluation des Copernic Summarizer und mögliche Einsatzfelder in der Fachinformation der DaimlerChrysler AG. Berichte aus der Wirtschaftsinformatik. Shaker Verlag GmbH, Aachen, Deutschland.

Heinz, Dr. Eberhard (2004): Interner Foliensatz der DaimlerChrysler AG. Stuttgart, Deutschland.

Herczeg, Michael (1994): Software-Ergonomie: Grundlagen der Mensch-Computer-Kommunikation. 1. Auflage. Addison-Wesley, Bonn, Deutschland.

Hofmann, Andreas (2000): Untersuchungen zur Beurteilung regionaler Technologiekompetenzen mit Patentstatistischen Analysen. Diplomarbeit an der Fachhochschule Stuttgart – Hochschule für Bibliotheks- und Informationswesen. Stuttgart, Jena, Deutschland.

Kamphusmann, Thomas (2002):Text-Mining: Eine praktische Marktübersicht. Symposion Publishing GmbH, Düsseldorf, Deutschland.

Krahl, Daniela; Windheuser, Ulrich; Zick, Friedrich-Karl (1998): Data Mining: Einsatz in der Praxis. Informatikzentrum der Sparkassen-Organisation GmbH. Addison Wesley Longman Verlag GmbH, Bonn, Deutschland.

Krause, Dirk (2002): Einsatzmöglichkeiten von Text-Mining zur Unterstützung von internetbasierten Ideenfindungsprozessen. In: Engelien, Martin; Homann, Jens (Hrsg.): Virtuelle Organisation und Neue Medien 2002: Workshop Gemeinschaften in Neuen Medien. Reihe Telekommunikation und Medienwirtschaft, Band 14. Josef Eul Verlag GmbH, Lohmar, Deutschland, S. 577-592.

Lorenz, Sascha (2001:) Text Mining – Methoden und Techniken. Diplomarbeit an der Technischen Universität Dresden. Dresden, Deutschland

Meier, Marco; Beckh, Michael (2000): Text Mining. Wirtschaftsinformatik, Band 2, Nr. 42, S. 165-167.

Neumann, Günter (2001): Informationsextraktion. In: Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Rolf; Langer, Hagen (Hrsg.): Computerlinguistik und Sprachtechnologie: eine Einführung. Spektrum Akademischer Verlag, Heidelberg, Deutschland, S. 448-455.

Panyr, Jiri (1986): Automatische Klassifikation und Information Retrieval: Anwendung und Entwicklung komplexer Verfahren in Information-Retrieval-Systemen und ihre Evaluierung. Sprache und Information, Band 12. Max Niemeyer Verlag, Tübingen, Deutschland.

Poetzsch, Dr. Eleonore (2001): Information Retrieval: Einführung in Grundlagen und Methoden. Materialien zur Information und Dokumentation, Band 5. 2. Auflage. Verlag für Berlin-Brandenburg GmbH, Potsdam, Deutschland.

Renz, Dr. Ingrid (1995): Offenlegungsschrift DE 19526264 A1. https://intra-patent.daimlerchrysler.com/patwww/download/netans/DE/19526264/A1/17612701.pdf. (Datum des Zugriffs: 12. März 2004).

Renz, Dr. Ingrid; Franke, Jürgen (2003): Text Mining. In: Franke, Jürgen; Nakhaeizadeh, Dr. Gholamreza; Renz, Dr. Ingrid (Hrsg.): Text Mining: Theoretical Aspects and Applications. Reihe Advances in Soft Computing. 1. Auflage. Physica-Verlag, Heidelberg, Deutschland, S. 1-20.

Rolker, Claudia (2002): Ein iteratives Information Retrieval Verfahren mit automatischer Suchmechanismenauswahl. Dissertation. Logos Verlag, Berlin, Deutschland.

Runkler, Thomas A. (2000): Information Mining : Methoden, Algorithmen und Anwendungen intelligenter Datenanalyse. Vieweg Verlag, Braunschweig, Deutschland.

Rupp, Chris (2001): Requirements-Engineering und -Management : professionelle, iterative Anforderungsanalyse für IT-Systeme. Sophist Group. Carl Hanser Verlag, München, Deutschland.

Scoreberlin GmbH (1999-2004): Artikel: Usability, genormte Qualität. http://www.scoreberlin.de/usability-artikel/usability-iso-norm/. (Datum des Zugriffs: 07. Juli 2004).

Wallmüller, Dr. Ernest (2001): Software-Qualitätsmanagement in der Praxis: Software-Qualität durch Führung und Verbesserung von Software-Prozessen. 2. Auflage. Carl Hanser Verlag München, Wien, Deutschland.

Wikimedia Foundation Inc. (2004): Wikipedia Hauptseite. http://de.wikipedia.org/wiki/Hauptseite. (Datum des Zugriffs: 07. Juli 2004).

Wild, Lothar; Wittman, Alfred (1990): Patentinformation und gewerbliche Schutzrechte. 3. Auflage. RKW-Verlag, Eschborn, Deutschland. Verlag TÜV Rheinland GmbH, Köln, Deutschland.

WIPO - World Intellectual Property Organization (2004): Frequently asked questions about the International Patent Classification (IPC). http://www.wipo.int/classifications/en/ipc/faq/ipcfaq-ver01.htm#P58_2444. (Datum des Zugriffs: 27. April 2004).

WIPO - World Intellectual Property Organization (2004): International Patent Classification. http://www.wipo.int/classifications/fulltext/new_ipc/index.htm. (Datum des Zugriffs: 05. Juli 2004).

Wurzer, Dr. Alexander J. (2003): Wettbewerbsvorteile durch Patentinformationen. 2. Auflage. Fachinformationszentrum Karlsruhe (FIZ Karlsruhe), Gesellschaft für wissenschaftlich-technische Information mbH, Eggenstein-Leopoldshafen, Deutschland.

Weiterführende Literatur

Baeza-Yates, Ricardo; Ribeiro-Neto, Berthier (1999): Modern Information Retrieval. ACM Press, New York, USA.

Becker, Peter (2004): Evaluierung von Retrievalsystemen. Vorlesungsunterlagen. http://www2.inf.fh-rhein-sieg.de/~pbecke2m/retrieval/eval1.pdf. (Datum des Zugriffs: 18. März 2004).

Brückner, Thomas (2001): Textklassifikation. In: Carstensen, Kai-Uwe; Ebert, Christian; Endriss, Cornelia; Jekat, Susanne; Klabunde, Rolf; Langer, Hagen (Hrsg.): Computerlinguistik und Sprachtechnologie: eine Einführung. Spektrum Akademischer Verlag, Heidelberg, Deutschland, S. 442-447.

Gentsch, Peter (1999): Data Mining und Text Mining als zentrale Technologien des Business Intelligence. IM - Fachzeitschrift für Information Management und Consulting, Saarbrücken, Band 14, Nr. 4, S. 23-28.

Gerstl, Dr. Peter; Hertweck, Dr. Matthias; Kuhn, Birgit (2001): Text Mining: Grundlagen, Verfahren und Anwendungen. In: Hildebrand, Knut (Hrsg.): HMD (Theorie und Praxis der Wirtschaftsinformatik), Business Intelligence. Band 222. dpunkt Verlag, Heidelberg, Deutschland, S. 38-48.

Heyer, Prof. Dr. Gerhard (2001): Text Mining - Ein Weg zur intelligenten Suche im Web: Grundlagen und Anwendungen. http://wortschatz.informatik.uni-leipzig.de/asv/publikationen/ErlangenTextMining.ppt. (Datum des Zugriffs: 08. März 2004).

Krause, Jürgen (1987): Inhaltserschließung von Massendaten: Zur Wirksamkeit informationslinguistischer Verfahren am Beispiel des Deutschen Patentinformationssystems. Linguistische Datenverarbeitung, Band 8. Georg Olms Verlag, Hildesheim, Deutschland.

Nakhaeizadeh, Dr. Gholamreza (Hrsg.) (1998): Data Mining: Theoretische Aspekte und Anwendungen. Beiträge zur Wirtschaftsinformatik, Band 27. Physica-Verlag, Heidelberg, Deutschland.

Reitzig, Markus (2002): Die Bewertung von Patentrechten: Eine theoretische und empirische Analyse aus Unternehmenssicht. Dissertation Universität München. Deutscher Universitäts-Verlag GmbH, Wiesbaden, Deutschland.

Stürmer, Uta (1995): Vergleichende Evaluierung verschiedener Interaktionsparadigmen für Information-Retrieval. GMD-Studien, Nr. 277. GMD – Forschungszentrum Informationstechnik GmbH, Sankt Augustin, Deutschland.

Weiss, Sholom M; Indurkhya, Nitin (1998): Predictive Data Mining: A Practical Guide. Morgan Kaufmann Publishers, Inc., San Francisco, USA.

Witten, Ian H; Eibe, Frank (2001): Data Mining: Praktische Werkzeuge und Techniken für das maschinelle Lernen. Carl Hanser Verlag München, Wien, Deutschland.

Erklärung

Hiermit erkläre ich, dass ich die vorliegende Diplomarbeit selbständig angefertigt habe. Es wurden nur die in der Arbeit ausdrücklich benannten Quellen und Hilfsmittel benutzt. Wörtlich oder sinngemäß übernommenes Gedankengut habe ich als solches kenntlich gemacht.

Ort, Datum

Unterschrift